Stages of Knowledge Discovery in Web Sites

Matthew Smith and Christophe Giraud-Carrier

Brigham Young University, Department of Computer Science, Provo, UT 84602, USA

Abstract

E-commerce greatly facilitates and enhances the level of interaction between a retailer and its customers, thus offering the potential for smarter marketing through thoughtful site design and analytics. This paper presents the three stages of knowledge discovery that move e-retailers from on-line catalog providers to finely-tuned, customer-centric service providers.

Keywords: E-commerce Intelligence, Web Mining, Clickstream Analysis, Personalization.

1. Introduction

As the Internet continues its expansion, most retailers have successfully added the Web to their other more traditional distribution channels. For too many companies, the story, unfortunately, ends there. That is, the Web channel is just that, another distribution channel, often consisting of little more than an on-line catalog tied to a secure electronic point of sale. Although valuable in its own right, such use of the Web falls far short of some of the unique possibilities it offers for intelligent marketing. Indeed, with nearly everything traceable and measurable, the Web is a marketer's dream come true.

Physical (i.e., brick-and-mortar) stores are static and generally customer-blind:

- The store layout and contents are the same for all customers and changes are often costly.
- Visits leave very little useful trace, other than perhaps, sale's data, generally limited to what was bought, when it was bought and by what method of payment.

On the other hand, on-line stores are dynamic and customer-aware:

- Layout and contents are easily modified and can even be tailored to individual visitors.
- Every visit automatically generates a trail of information on the customer's experience (e.g., how long she stayed, what she looked at, whether she bought or not, etc.), and possibly on the customer herself.

Unfortunately, due to lack of knowledge, Web capabilities and data are too often under-exploited in E-commerce. The full benefit of the emerging and growing Web channel belongs to those who leverage the rich information it provides [7, 10]. To that end, this paper describes three stages of knowledge discovery in Web sites: (1) clickstream analysis, (2) advanced Web mining, and (3) personalization. Moving through these stages requires increasing sophistication, but also produces increasing return-on-investment.

2. Clickstream Analysis

The readiest source of information or data to mine on a Web site is Web usage data. Web servers are generally configured so as to store such data, also known as clickstream data, in Web server log files. If not, they can be set up to do so easily and quickly.

Web server logs grow as visitors interact with the Web site. For each visitor, the log contains basic identifying information (e.g., originating IP address) and time-stamped entries for all visited pages, from the time the visitor first reaches the site to the time she leaves it. The amount of data found in Web server logs is enormous and clearly evades direct human interpretation. However, log analysis tools (e.g., AWStats [2], The Webalizer [3], SiteCatalyst [15], ClickTracks [6], NetTracker [17], etc.) can be used to summarize it, allowing marketers to gain some knowledge about their e-commerce activities. In particular, the following questions can be answered:

- Where do most visitors come from?
- How many of visitors come from a direct link or bookmark vs. a search engine vs. a partner Web site (if any)?
- What 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 they visit on average?
- Which pages are most popular?
- When do they leave the site?
- Which pages do visitors commonly leave the Web site from?

Hence, the result of analyzing Web server logs allows reporting on unique visitors, hits, visit duration,

This work is supported in part by the BYU Bookstore and Overstock.com.

visitor’s paths through the site, visitors' host and domain, search engine referrals, robot or crawler visits, visitors' browser and operating system (OS), and more. This information in turn helps e-retailers better appreciate the dynamics of customers’ interactions with their on-line offerings, and begins to inform such decisions as which referrer to invest in, which pages to remove or replace, which pages to improve, etc.

For instance, assume that clickstream analysis shows that a substantial number of visitors are accessing content several clicks deep into the Web site. Then it might be wise to make that content more accessible so visitors can find it faster without having to click so many times. Methods for doing this include maintaining “Top Sellers,” “Most Wished for Items,” or “Most Popular Items” pages.

However, there is a limit to the knowledge that may be extracted from Web server logs. In general, transactional data and order information are not stored directly in standard server logs. Yet, these are essential in discovering patterns of buying and non-buying customers. Linking clickstream data to (transactional) order information, together with thoughtful design, allow marketers to take further control of their e-Commerce activity through more advanced Web mining.

3. Advanced Web Mining

As mentioned earlier, e-commerce activities have the potential for richer data footprints than their physical counterparts, thus providing greater opportunities for intelligent e-business through Web data mining. However, as with all other forms of successful data mining, Web mining cannot be an afterthought; it requires planning, which in turn translates into informed design decisions.

With adequate design, a host of new business-relevant questions may be answered, such as:

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

As an example, assume an e-retailer wishes to run a special offer and measure its impact. Then, a mechanism capable of capturing the supporting data must be set in place. In the on-line world, this may be accomplished simply and cost-effectively by adding a special ad somewhere on the Web site and tracking the visitors who click on it (typically done using a distinguisher in the URL). At any time, one can easily compute the offer’s targeting success rate by taking the ratio of those who click and convert (i.e., buy the special offer) to those who only click. Although not completely accurate, return-on-investment may also be easily computed as the experienced increased in revenue following the offer versus its cost.[1] Extending to multiple offers, this functionality provides a method to determine which offers are most successful, thus informing marketing decisions. Interestingly, referral programs can be implemented and evaluated using the same approach.

Advanced Web mining requires at least correlating both visitors’ clickstreams and transaction information. However, in order to answer the more customer-related questions, one must also be able to associate specific behavior (i.e., Web usage and buying patterns) to specific customers. Again, this impacts site design.

The only way to accurately identify visitors is by having them sign in (e.g., log on to the Web site), thus creating a session. This is particularly important as visitors may be accessing the Web site from multiple computers in varied locations such as home, work, or grandma's house. Log-only information would in this case give the impression of three distinct visitors. When a visitor signs in, however, he or she can be uniquely identified and information can be associated to them, such as interests, preferences, profile (e.g., gender, address, age, etc.), purchasing habits and clickstream history. This type of session tracking is commonly implemented with “cookies” (a mechanism for storing data in the remote visitor's browser) or by propagating a unique session id in the URL. With this additional data, questions of segmentation and profitability can readily be addressed.

Note, however, that in many instances signing in to a Web site may be a deterrent, mostly due to concerns with privacy and spamming (e.g., see [5, 18]). A clear privacy statement and special privileges for signed-in users are often sufficient incentive. Both customers and e-retailers are learning that the benefit of collecting more information about visitors exceeds the cost.

[1] Accurate measurement would require proper experimental setting with a control group to factor out revenue increases that may result from something independent of the offer. In practice, it is often assumed that other things remain equal and the direct measurement is used as a reasonable approximation for decision purposes.

With advanced Web mining, e-retailers take control of their marketing activities and learn what works well, which customers deserve extra attention, and what should be offered to whom. Although much of this additional knowledge may be leveraged within the context of a static Web site, its real benefit comes as e-retailers recognize that they can engage more fully with their customers by offering the right thing to the right customer *at the right time*.

4. Personalization

Every marketer's dream is to know enough about customers so as to tailor her offer to each individually, in terms of both products and prices. Even when the needed knowledge is available, this is nearly impossible in a traditional store. Internet technology, on the other hand, makes it possible to adapt layout, contents and services offered "to the needs of a particular user or a set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the Web site" [8].

Furthermore, the Web channel taps into an unusually diverse and large customer pool. Web customers are not restricted by physical geography; they can come from all over the world and exhibit widely different demographic and socio-economic characteristics. This abundance of data offers a unique opportunity for personalization.

Unlike others, who distinguish several forms of personalization, such as content-based filtering or collaborative filtering (e.g., see [8]), we explicitly view personalization as the final stage of knowledge discovery in e-commerce, where data, made available through careful site design, is mined and translated into finely-honed marketing actions:

- Show product *P* first to customer *C*.
- Offer a discounted price on bundle *B* to customer *C*.
- Suggest services and products that customer *C* is likely to be interested in.
- Provide timely chat or co-browsing to most valuable customers [4].

As a simple example, consider the prevalent notion of "gold" or high-value customers, consisting of that (generally small) group of customers which generate most of the revenue and profit [11]. Here, personalization makes it possible for both high-value and low-value customers to be precisely identified and treated accordingly during each visit. Customer appreciation programs and benefits can be used to retain high-value customers, while low-value customers can be enticed to become high-value customers through special offers.

Interestingly, with dynamic web design, one may use both information previously obtained and data provided in real-time (i.e., as the visitor interacts with the site) to tailor the exchange between the parties. Some of the largest e-businesses have had enormous success personalizing Web content. We mention a few here as an illustration.

Google presents relevant advertisements based on keywords in which a visitor is interested [9]. Overture, now Search Marketing [20], a Yahoo! company, provides a similar service for sites including AltaVista [1], MSN [13], and Yahoo! [19]. Amazon.com uses collaborative filtering to recommend products to users on-the-fly. In this case, order information is leveraged to identify clusters of users with similar purchasing habits, the underlying assumption being that people with similar buying behavior are very likely to have similar interests [8, 12]. Yahoo!'s LAUNCHcast also uses collaborative filtering to recommend music to listeners. In this scenario, however, music ratings are elicited from visitors and used to identify clusters of users with similar musical tastes [21].

Although quite useful in their own right, these implementations only scratch the surface of what is achievable with personalization in e-commerce. Our own research focuses in using richer sets of data to build more versatile models for personalization, to support such activity as cross-selling and up-selling. Thanks to the Internet's flexibility, e-commerce is probably the closest one can hope to get to 1-to-1 marketing.

Ideally, personalization benefits both the customer and the e-retailer. Successful recommendations benefit customers by readily providing them with items most likely to be of interest, sometimes even introducing items that they were previously unaware of. In fact, not only are the most relevant products delivered but the transaction itself requires less time [16], thus improving overall customer satisfaction. In turn, more visitors will convert (i.e., buy what is suggested to them) and exiting customers will return (i.e., build loyalty), so that the e-retailer sees increased revenue and profit at a relatively small cost.

5. Conclusion

In this paper, we have briefly reviewed the value of planning for knowledge discovery in the design of e-commerce applications. Three incremental stages of knowledge discovery have been presented and illustrated with simple examples.

The sequence of stages has been designed to bring increasing return-on-investment at each stage and to help e-retailers get closer to optimal use of the Web channel. Indeed, the unique nature of the on-line world makes achievable the double objective of maximizing the customer's experience and maximizing revenues.

We have focused here on enhancing e-commerce through personalization, in particular by matching offers to interests (i.e., supply and demand) in an efficient way, via Web data mining. There are many other data-independent things e-retailers can do to enhance the user experience, such as offering social interactions (e.g., user forums) or providing virtual versions of physical stores (e.g., displays, lighting, music) [14].

6. References

[1] Alta Vista, 2005. (www.altavista.com).

[2] AWStats. *Logfile Analyzer 6.4*, 2005. (awstats.sourceforge.net/docs/index.html).

[3] Barrett, B.L. *The Webalizer*, 2005. (www.mrunix.net/webalizer).

[4] Beagle Research Group. *Turning Browsers into Buyers with Value-Based Routing: Methodology Enhanced e-Commerce*, White Paper, 2004. (www.beagleresearch.com/DownloadPDFfiles/ProficientFINAL.pdf).

[5] Center for Democracy and Technology. *CDT's Guide to Online Privacy*, 1998. (www.cdt.org/privacy/guide/introduction).

[6] ClickTracks, 2005. (www.clicktracks.com).

[7] Edelstein, H.A. Pan For Gold In The Clickstream. *Information Week*, March 2001. (www.informationweek.com/828/prmining.htm).

[8] Eirinaki, M. and Vazirgiannis, M. Web Mining for Web Personalization. *ACM Transactions on Internet Technology*, Vol. 3, No. 1, 1-27, 2003. (www.db-net.aueb.gr/magda/papers/TOIT-webmining_survey.pdf).

[9] Google. *Advertising Programs*, 2005. (www.google.com/ads).

[10] Greening, D.R. Data Mining on the Web. *Web Techniques*, January 2000. (www.newarchitectmag.com/archives/2000/01/greening).

[11] Hughes, A. 24 Key Database Marketing Techniques. *DM News*, January 2005. (www.dmnews.com/cgi-bin/artprevbot.cgi?article_id=31602).

[12] Linden, G., Smith, B. and York, J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, Industry Report, January/February 2003. (dsonline.computer.org/0301/d/w1lind.htm).

[13] Microsoft. *MSN Search*, 2005. (search.msn.com).

[14] Oberbeck, S. Internet shopping is big business, but more can be done by e-retailers. *The Salt Lake Tribune*, December 2004.

[15] Omniture, Inc. *SiteCatalyst 11.1*, 2005. (www.omniture.com/s2/sitecatalyst.html).

[16] Personalization Consortium. *Personalization Information*, 2005. (www.personalization.org/personalization.html).

[17] Sane Solutions LLC. *NetTracker Web Analytics Solutions*, 2005. (www.sane.com).

[18] World Wide Web Consortium (W3C). *Platform for Privacy Preferences (P3P) Project*, 2004. (www.w3.org/P3P).

[19] Yahoo! Inc., 2005. (www.yahoo.com).

[20] Yahoo! Inc. *Search Marketing*, 2005. (searchmarketing.yahoo.com).

[21] Yahoo! Inc. *LAUNCHcast*, 2005. (launch.yahoo.com).