Driving e-Commerce Profitability From Online and Offline Data

A Torrent Systems White Paper

TORRENT®

Torrent Systems, Inc.
Five Cambridge Center
Cambridge, MA 02142
617/354-8484
617/354-6767 FAX

For more information, e-mail Torrent at info@torrent.com

Executive Summary

Web traffic is no longer enough. In today's Internet climate, investors are demanding that *all* e-commerce sites generate significant revenue, even profit. Unfortunately, according to a recent article published in *The Wall Street Journal*, of the 30,000 e-commerce sites identified by shop.org, the Internet retailing trade association, only 3.3 percent generate sales in excess of $500,000[1]. What's needed to drive e-commerce profits is the ability to convert casual browsers into loyal customers.

Maximizing the conversion rate requires increasing each visitor's likelihood of commitment (through purchases or registration) by first understanding visitor behaviors and then personalizing in real time the content delivered to each Web site visitor. To this end, many Web reporting and Web personalization software tools have emerged in the market. These first-generation tools have three major shortcomings: they offer little function for integrating valuable offline data sources; they can't process and analyze the large volumes of clickstream data in timely fashion; and they rely on simplistic data preparation, reporting and analysis techniques.

Prevailing in today's hypercompetitive e-marketplace requires the application of sophisticated, scalable data-mining techniques to large volumes of data–both online clickstream or Web usage data and offline data, such as history, demographic, and transaction data from a data warehouse or data mart–for the purpose of optimizing Internet offers and tailoring site content in real time. The keys to deploying a successful Web usage mining system include:

- a framework that allows the integration of disparate data processes and data formats, to leverage best-of-breed processes already in place and those to come
- a parallel, scalable infrastructure, to support "the need for speed": real-time processing and analysis of the massive quantities of clickstream data flowing through and stored by today's Web sites
- a rich offering of scalable, in-depth data mining tools for analysis of Web usage patterns and user preferences and for organizing patterns into predictive models of user behaviors.

Most important, the effectiveness of each of these attributes depends heavily upon the other two. For example, for a framework to allow integration of disparate data, it must be able to scale to accommodate new formats and new sources. The same

1. WSJ, April 5, 2000

holds true for analysis: the system must permit the addition of new functions without becoming bogged down. In short, the three keys are inseparable.

1. Online Business Changes Retailing Rules and Risks

The customer's experience of shopping online contributes greatly to the hypercompetitive nature of the e-marketplace. Online visitors can jump sites instantly and effortlessly; competition is only one click away. For instance, research by Jupiter Communications indicates that 53 percent of online buyers visit at least three sites before making a purchase. This seamless comparison places Web businesses at a disadvantage compared with brick-and-mortar businesses. Walking into a store involves inertia, if not loyalty: it takes effort to switch stores. A customer in a physical Barnes & Noble's bookstore will need strong incentive to walk across the mall (let alone, drive) to Borders. The same customer browsing through bn.com will think nothing of clicking to amazon.com. A customer's mouse is always faster than his or her feet.

Applying to brick-and-mortars the jargon of e-business, one can say that brick-and-mortars have over e-commerce the advantage of inherent *stickiness*. The laws of physics make customers "stick" longer in physical stores than in virtual ones.

Fortunately, there is a huge flip side in favor of online business. E-commerce vendors can achieve what brick-and-mortars can only dream of: in effect, the ability to attach a homing device to all shoppers entering a store, track their paths through the aisles, determine the displays and promotions that catch their eye, monitor how they add items to and remove them from their shopping carts, etc. Web usage mining does just that.

The aggressive competition in the e-marketplace makes the proper application of Web usage mining all the more critical to securing competitive advantage. According to research by GartnerGroup, of the 110 million individuals online, 79 percent use the Internet to research products and purchases, while only 41 percent actually place orders online.[2] To prevail in today's e-marketplace, Web sites will need to convert these browsers into buyers.

2. GartnerGroup, *Buy-To-Browse Ratios for Internet Shoppers*, Feb. 2000

Specifically, they will need to:

- attract more first-time visitors to the Web site
- keep visitors on the site longer (increase site stickiness)
- turn first-time visitors into repeat visitors
- compel visitors to register
- generate higher click-throughs to targeted ads
- convert lookers into first-time customers
- convert first-time customers into repeat, profitable customers.

2. Web Usage Mining Is a Prerequisite of Personalization

Certainly, personalization is not a new retailing concept. Ever since people first traded goods and services, businesses have understood that identifying the behaviors and preferences of customers is the key to responding effectively and competitively. What is truly new, however, is the irreversible impact of the Internet on the way vendors gather business insight and act upon it. The rich customer data that flows into a Web site every day allows businesses to understand—and serve—their customers not as demographic groups, but as individuals.

On the Web, personalization acts upon the models and profiles obtained through analysis and dynamically serves customized content (links, products, recommendations, ads, services, pricing offers, etc.) to individual visitors. It is important to note that successful personalization demands a rigorous preparatory analysis phase that processes both clickstream and offline data, such as demographic, history, preferences, and other relevant information. Given the multiple sources of offline customer data and the sheer volume of both offline and online data, Web usage mining presents a formidable challenge. But it also offers great rewards: because user clicks represent real interactions between businesses and customers, Web usage mining offers significant promise for developing much richer profiles of user preferences and behaviors and for large-scale targeting of personalized content to individuals.

Indeed, the Web serves as a direct channel between vendors of products and services and data that reveals the preferences of prospects and customers. Every time a visitor clicks on a Web site, he or she is indicating a choice or preference. The abil-

ity to collect detailed data at the granular level of individual mouse clicks and then perform sophisticated analysis of large collections of these individual Web clicks–augmented, whenever possible, by transaction and demographic data–provides vendors with a tremendous opportunity for personalizing the dynamic Web experience.

To be exploited successfully, the mass of collected Web usage data must first be *mined* and then transformed into information that can be leveraged for real-time personalization. For example, Web usage mining helps build models for understanding and predicting user behaviors by:

- determining fruitful categorizations of visitor usages
- uncovering predictive association rules, e.g., between site organization and navigation patterns
- isolating affinities between accessed pages.

Unfortunately, first-generation Web reporting and personalization tools cannot present this analysis with the speed and sophistication required by today's user. The current tools force businesses to sacrifice analytical depth, speed of response, or both. At the heart of these problems lies a handful of technological deficiencies. They are as follows.

- inability to smoothly integrate clickstream or Web usage data with offline data sources
- inability to handle massive volumes of data
- the faster the response, the less sophisticated the nature of the analytical algorithms used.

3. The Three Keys to Prevailing in the Internet Economy

Experience in large-scale personalization strongly indicates that to realize the goals of the previous section, one must bypass the technical limitations of the current crop of Web reporting and personalization tools to achieve:

- integration of rich sources of customer information from offline sources, and existing (tried-and-true) processes
- scalability to handle massive data volumes
- analytic depth in Web usage mining.

Indeed, to reorganize their Web sites to maximize the return on investment for each visitor, portals, content providers, e-commerce sites, and other Web players must access and leverage intelligently every bit of information contained in clickstream and offline data.

3.1 Integration Is the First Key

By integrating disparate application processes and data-enriching offline sources, technology builds the rich user profiles required for e-marketing success.

To fully develop customer profiles, offline data warehouses now are adding clickstream data into their stores, while Internet warehouses are integrating transaction and demographic data. As a result of this online and offline merger, the customer data available is becoming increasingly rich, but integrating the disparate sources of this data, and mining the aggregate volumes in a timely, orderly fashion is becoming equally challenging.

But absolutely necessary. To build a complete model of the user, Web usage mining *must* compare clickstream with offline data, such as demographic data, archived information on past business interactions with a particular visitor (or visitor's classified group), call-center transactions, purchase history, etc. In practical terms, Web usage mining must handle the integration of offline data with:

- e-business analytic tools
- various e-business RDBMS's
- e-business catalogs of products and services
- e-business customer service/support departments
- best-of-breed applications.

Unfortunately, most existing reporting and analytic systems cannot effectively integrate these disparate data sources and processes. Opting to do without this capability results in a fragmented view of the customers that inhibits the ability to cross-sell, up-sell, enhance customer loyalty, convert visitors to customers, and other tasks identified as critical to prevailing in the aggressive e-marketplace.

Another major benefit to technology that facilitates integration is that it provides the flexibility to customize applications for competitive edge. The canned reports provided by existing Web reporting tools do not enable major portals, e-commerce

sites, content providers, and other major Internet players to develop insights that differentiate them from the competition.

3.2 Scalability Is the Second Key

The more data you feed into your analysis, the greater the likelihood of discovering hidden patterns that lead to better e-marketing; large data volumes require scalable processing.

High-quality modeling requires large data samples with lower estimation errors and lower variance. As data volumes and sample sizes grow and the window of time for analysis shrinks, the greater the requirement for scalable processing. The problem is that most application processes and Web reporting and analytical tools are designed to run on only one processor, regardless of the number of processors available to work in concert on any one problem. The amount of data quickly overwhelms the power of any one processor. Unfortunately, converting these sequential processes and tools to run in parallel is either impossible or exceedingly difficult.

What is required is an environment that enables programmers to develop Web usage mining systems that run in parallel to handle the enormous volumes of data in the time frame required by the Internet economy. This environment must remove the complexities of parallel programming from the development process, easing the construction of in-house systems that offer a unique competitive edge. In addition, this environment must manage the parallel execution of these complex systems against massive data volumes, handling the partitioning and re-partitioning of data across myriad application processes, the replication of application logic, parallel communication with the RDBMS, and parallel scalable analysis. These capabilities, which have proved indispensable in the data warehousing arena, are no less critical to the survival of e-businesses.

3.3 In-Depth Web Usage Mining Is the Third Key

In-depth Web usage mining offers the greatest potential for e-marketers to discover and leverage models of visitor behavior.

According to Jupiter Communications, the size of the Web will double every three months, with usage increasing at a similar exponential rate. Independently, the cost of storing data has also fallen precipitously. Today, businesses save data they previously might have discarded. As it became increasingly cost-effective to do so, e-

merchants started to save *all* the clickstreams from their sites at the granularity of individual clicks.

As a consequence of the growth of archived Web-traffic data, today's e-merchants sit atop a gold mine—the accumulated data deposited by their visitors. This gold mine, however, remains largely unexploited because, until now, the technology has not existed that offers both in-depth analysis of Web usage data and the ability to handle massive volumes of clickstream and offline data in a timely fashion.

Web usage mining strives to make sense of this information by analyzing the clickstream information collected in Web server logs, referrer logs, user registration forms, and purchase data. Web usage mining helps e-marketers determine the lifetime value of customers, cross-market products, up-sell, rate the effectiveness of promotions, and convert visitors to customers by tailoring Web content to individual visitors.

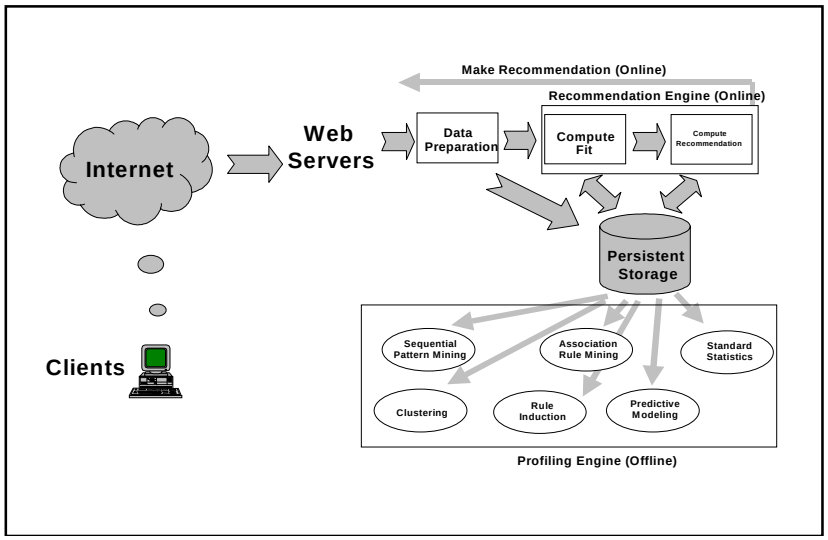

FIGURE 1. Influence of data mining on personalized online experience.

Data mining typically involves three steps: preparing and gathering usable data, mining that data to build models of behavior, and finally, deploying these models for personalization.

Preparing Web Data for Analysis

To unlock the secrets contained in Web data, companies first must reconcile and clean the varied and disparate manners in which this data is presented. In many respects, this process is similar to the Extract, Transform, and Load (ETL) process used in traditional offline data warehousing and data mining. But on the Web, the ETL stage presents a number of new and daunting tasks. These are as follows.

- **Gathering and Consolidating User Data:** To handle increasing traffic demands, many of today's large, advanced Web sites have adopted load-balancing structures, whereby traffic is constantly distributed across all available servers. While this transfer is seamless to the end user, records of user interactions with the site are anything but. As a result of the load balancing, user data is spread through the logs of all employed servers, which in a typical site can number anywhere from a few to well over a hundred. Thus, in order to track a user's interaction with the site (i.e., establish the clickstream), it is necessary to compile the logs from all servers and then reorder the data, grouping by user and ordering by the time at which files were requested.

  The already-difficult task of gathering and sorting data by user can be complicated further by the manner in which sites identify their users. Currently, Web sites can identify users by three methods: cookies (small text files downloaded from the site and imbedded in the user's browser), URL rewriting (the process of inserting tags into the URL's requested by a particular user), and a user's IP addresses. For tracking and sorting user data, cookies are by far the most dependable–and fruitful–method of identification as they establish a positive, enduring ID. URL rewriting is functional as well, but less effective than cookies; while it tags the URL's requested by a user, it does so only on a session-by-session basis and thus cannot track single users over multiple sessions. IP addresses, on the other hand, provide no such reliable tags. A single IP address actually can represent a multitude of unrelated users (sometimes called the "AOL effect"), which prevents accurate tagging and sorting. Conversely, load-balancing applications and privacy tools from proxy servers and fire walls will often attribute multiple IPs to a single user over the course of a single session.

- **Data Cleansing:** In addition to pertinent user data, the Web server log files contain volumes of redundant or irrelevant information, all of which must be filtered out prior to analysis. Such extraneous information may include out-of-range values and entry errors, error codes, request methods other than "GET," image files, information requested by "spiders," and post events.
- **Content Labeling:** While gathering and consolidating user data may render the stream of URLs requested by a user, without knowing the exact content served, it offers little indication of usage patterns or behavior. In the days of static Web sites, base URLs may have been a good indication of the content served to a user, but with the trend today toward dynamic content, the permutations of content served—and by association the URLs themselves—can be nearly limitless. Thus, to accurately identify the precise content requested by an individual user, the query string of each URL must be searched for known tags and tag combinations.
- **Data Integration:** With the proliferation of Application Service Providers (ASP's) comes the challenge of data integration. While the ASP's may serve ads (DoubleClick, AdForce), affiliate marketing programs (BeFree), or search functions (Inktomi) to a Web site, all records of customer interactions reside on the ASP servers, not the host Web site. Thus, to get an understanding, for example, of what ads brought customers to a Web site and ultimately what promotions attracted their attention, it is necessary to integrate ASP customer data with data derived from host server logs. Potentially, this integration may require that ASP data be cleaned and sorted to remove extraneous information, and reformatted to match current databases. Also, the ASP data must be correlated with Web data from the host site.

Sophisticated Data Mining Methods

Data mining methods come in two flavors: descriptive and predictive (often associated with “unsupervised” and “supervised” learning, respectively). Descriptive and predictive methods are typically used in sequence. For example, Web usage mining employs clustering to unearth inherent classes, and then applies the obtained characterization to distribute events into the classes.

Descriptive methods include:

- Clustering: organizes visitors into similar and meaningful but often hidden groupings. Advantage: requires no *a priori* knowledge on nature of groupings; groups are an output

- Pattern Discovery: finds descriptive patterns in visitor behaviors (e.g., effects on site organization and topology on traversal patterns, hidden characteristics of visitors inclined to register or buy)
- Dimensionality Reduction: simplifies modeling by reducing the number of independent variables.

Predictive methods include:

- Prediction: estimates likelihood of particular visitors to exhibit patterns found by pattern discovery, e.g., register or buy
- Classification: organizes visitors into predefined categorical groups, or "classes." Classification is a kind of "supervised clustering" where the classes are known inputs, rather than outputs (as in regular, unsupervised clustering). The classes can be the obvious groups (visitors who click on ads, who buy) or less evident classes obtained typically as an output of clustering
- Estimation/Regression: same as classification but with numerical data (e.g., age, income, monthly credit card usage). Requires knowledge of history
- Affinity: finds relationships between visitors and products
- Optimization: maximizes value factors like stickiness, likelihood of registration, and transaction
- Time Series: Intelligent extrapolation of future values for a few selected variables based on the analysis of their past values
- Rule Induction: Analysis of data that induces predictions of the form of "If-Then-Else" rules.

These sophisticated data mining methods are implemented with a variety of algorithms:

- Neural Networks and their variants: probabilistic neural networks, radial basis function networks, neural trees, Kohonen networks
- Simulated annealing
- Regressions
- Principal component analysis
- Sensitivity analysis

Shortcomings of Current Reporting and Personalization Tools

Most existing Web reporting tools provide primitive mechanisms for reporting user activity. Basically, they distribute clicks into bins and construct histograms. While these tools enable vendors to determine the number of accesses to individual files, the times of visits, and URLs of users, they typically provide little analysis of data relationships among accessed files. This analysis is essential to leveraging fully the data gathered in daily transactions. Some of the more sophisticated reporting tools available today are capable of fairly in-depth analysis, but unfortunately they are not fast enough to enable real-time Web personalization.

As for the real-time "personalization" tools themselves, their limitation is in shallowness of analysis: they typically limit themselves to collaborative filtering techniques and other primitive, *product*-centric correlations. This shortcoming does not result from the failure to recognize the importance of analytical depth to competitive advantage. It is a consequence of a technical barrier, namely the difficulty of performing sophisticated data mining on huge volumes of data in a timely fashion. Most of the collaborative filtering and similar tools actually discard the granular clickstream data containing rich information about user choices and preferences in order to complete processing in the time needed.

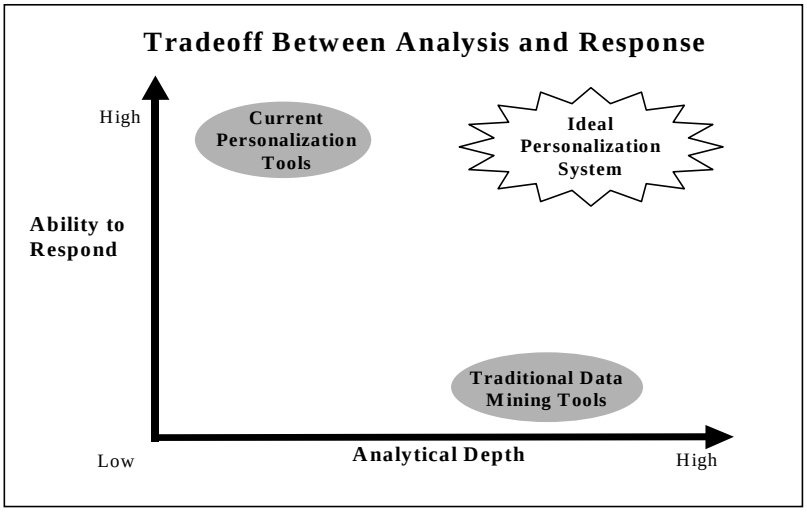

FIGURE 2. Current personalization and data mining tools can be either rich or responsive, but not both.

A wide gap is rapidly developing between the function delivered by first-generation Web reporting and Web "personalization" tools and the needs of the major Internet players (portals, media companies, application service providers or ASP's, and large e-commerce sites). These businesses need to implement sophisticated systems that can analyze massive volumes of Web clicks, transaction data, and data-enriching offline information in the time frame required for real-time personalization.

4. Implementing the Three Keys: Orchestrate

Torrent Systems' Orchestrate parallel software infrastructure enables major Internet players—portals, click-and-mortar businesses, e-commerce sites, content providers, and ASPs—to meet the requirements of Web usage mining for a decided competitive advantage. Orchestrate's architecture is uniquely optimized to provide:

- a framework that allows the seamless integration of disparate data formats and application processes, to leverage best-of-breed processes already in place, and those to come
- a scalable, parallel infrastructure, to support "the need for speed": real-time processing and analysis of the massive quantities of clickstream and offline data stored by today's Web sites
- a rich offering of scalable, sophisticated data mining tools for analysis of Web usage patterns and user behavior for predictive modeling.

The practice of field-deployed large-scale Web usage mining shows, however, that integration, scalability, and in-depth analytical tools alone are not enough. Orchestrate provides the needed additional features which include portability, component reuse, and ease-of-use.

4.1 Portability

Integration comes with a secondary but crucial requirement: portability to various parallel architectures and system sizes. Data warehouse and data mart systems often evolve in both size and shape. Changes in size are the most common; exponential growth of the Web and Web traffic makes frequent increases in scalability necessary. To vary the number of processors and adapt to changes in operating system (from MPP, to SMP clusters and SMP), all Orchestrate requires is editing a configuration file that is read at run time, thus removing a re-compilation step from the process.

4.2 Component Reusability and Robustness

Integration in Orchestrate leverages a remarkable feature of the modern RDBMS: the ability to run applications against different tables and against data with differing record structures, without requiring any program changes. As a result, applications are reusable and robust. The RDBMS tracks and manages record schema information, running checks prior to program execution to detect any inconsistencies. Without such support, any change to the record format would require changes to the application. Web usage mining must be designed to support the same level of component reusability and robustness for portions of Web mining applications that run *outside* of the databases, to ensure overall application integrity and reusability.

4.3 Ease of use

The need to combine scalability and integration is widely recognized. In practice, however, users under time constraints often forego this benefit in view of the difficulties of managing the complexities of programming multiprocessor systems. Orchestrate is designed to make the management of parallel resources transparent to the user; it offers C++, UNIX shell, and graphic interfaces.

In the figure on the following page, Jupiter Communications clearly places Torrent Systems “above the crowd.” In the current vendor landscape, Torrent holds a unique position characterized by the ability to process volumes in excess of 30 million daily page views. The current trends inescapably show that volumes of this size will become commonplace soon.

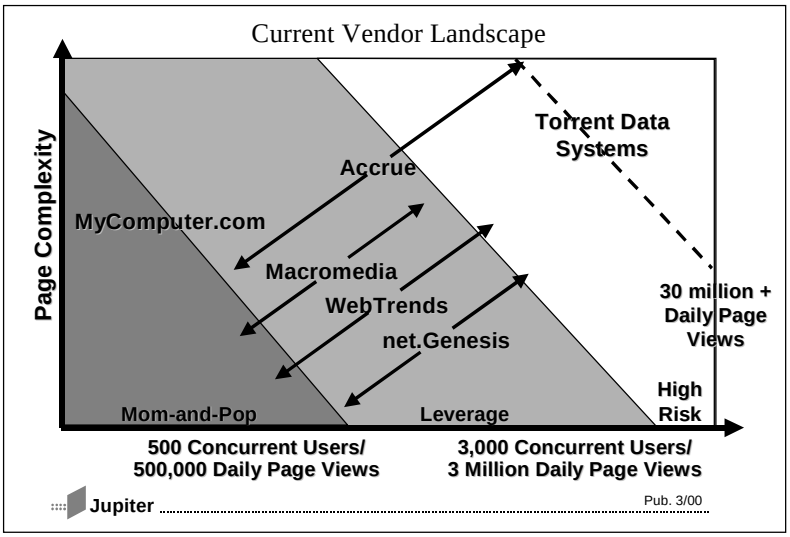

FIGURE 3. Jupiter Communications places Torrent Systems atop current vendor landscape.

5. What Makes the Three Keys Inseparable

5.1 Connection Between Scaling and Integration

Web analysis typically relies on accessing large amounts of data archived in data warehouses and data marts. Extracting this data and then transforming it and integrating it for unified analysis is best accomplished by parallel processing and data movement.

5.2 Connection Between Scaling and Analytic Depth

The most sophisticated analytic models are often the most time-consuming to train, a characteristic that has limited their applicability. However, parallel training facili-

tates the use of these powerful nonlinear models in situations in which they would otherwise be infeasible. The need for parallel processing is exacerbated by the fact that analytic modeling is not only CPU-hungry, but in practice it also must be repeated using multiple combinations of methods and algorithms.

Section 3.3 lists data mining methods and algorithms, begging the question: How to match methods and algorithms? The short answer is, you don't. The demands of Web analytics are too complex to be satisfied by a simple matching of methods and algorithms. Web analytics really requires a *process*, as does, for that matter, all data analysis when applied seriously. Academic research often refers to the algorithm-to-method matching approach as the "cookbook fallacy." [3]

There are several reasons why the algorithm-to-method approach is naïve:

- most methods overlap with other methods and most algorithms overlap with other algorithms
- a single method can be implemented by several algorithms, e.g., predictive methods can be implemented using KD Trees and most flavors of Neural Networks, and clustering can be implemented using Kohonen nets as well as neural trees
- while the business goals may be articulated clearly, it is generally very difficult to achieve satisfactory precision in *stating* the problem at the data level. Web analysis is indeed an interactive process: uncovering patterns leads one to refine old questions and ask new, more pointed ones. Furthermore, prototyping does not work well: smaller sample data will require different tools (and answer different questions) than the full-size problem. In other words, a particular tool, or even general approach can match problems at the prototype level but breaks down in production: hence the need for an iterative process, looping over different combinations of analytic methods and algorithms.

There is another instructive interplay between scaling and analytics that hinges on the relative sizes of the variance and bias errors which together contribute the total mean square error of a statistical estimator. Variance measures how widely the distribution of data points is scattered around the mean and it decreases proportional to the inverse square root of the sample size. Bias is an essential limitation of the model resulting from assumptions made in building it, such as assuming linear

3. See, e.g., M. Berthold and D. Hand (Ed.), *Intelligent Data Analysis* (Springer, 1999).

behavior in linear regression modeling. With relatively small samples, the systematic bias errors are hidden by the relatively large variance error. With larger amounts of data, on the other hand, the variance can become so small as to reveal the inherent bias. At some threshold in the size of the data, it becomes advisable to switch from biased methods (such as regression) to unbiased ones (such as implemented by neural networks).

Let us close with the connection between analytic depth and a variant of parallelism: pipelining. Pipelining accelerates the processing of a stream of clicks (or of database records) by allowing one operation to pass along a piece of data (click representation or record) as soon as it processes it to another operator downstream to it, rather than waiting for the processing of the entire stream to complete. Data flow architectures support pipeline parallelism and constitute a natural fit to the typical data mining practice where, as mentioned in Section 3.3, different tools are typically used back to back, following this example:

[Clustering] –> [Classification] –> [Optimization]

Orchestrate is based on a dataflow architecture and it fully supports pipeline parallelism.

5.3 Connection Between Integration and Analytic Depth

In many important Web reporting and personalization situations, even a comprehensive panoply of advanced data mining methods does not suffice. These methods must be supplemented by the more traditional approaches such as SQL queries and multidimensional analysis. The difficulty is these traditional approaches demand access to offline data, usually from large data warehouses. As a result, integrating these data stores into a hybrid system usually requires an extract, transform, and load (ETL) process similar to that used in traditional data warehousing.

Conclusions: The Benefits of Accessing Both In-Depth Analysis and Responsiveness

Today, the best Web reporting and personalization tools offer either in-depth analysis or swift response, but not both. The benefits of combining both features are enormous. Together, depth and speed enable e-commerce sites, portals, content providers, and other Web key players to maximize the return on investment of each

visitor by optimizing their Web sites and by delivering customized content to individual visitors. Thus, Web players achieve unprecedented high-velocity e-marketing quality.

Quality Web personalization, the kind that leads to sustainable business with profitable customers, is demanding. It requires rigorous data preparation, integration of online and offline sources, data analysis at the granularity of individual clicks (a source of information previously untapped), scalability to handle growing data volumes, and in-depth use of Web usage mining.

The keys to achieve these are a comprehensive offering of Web sophisticated analytic tools, integration, and scalability. Torrent Systems' Orchestrate is a parallel software infrastructure uniquely optimized to implement these three keys.