A Survey of Behavioral Targeting and its Technologies and Issues

Fiona Luong
Stephen Slaughter
Collette Spence

INFO606
Dr. Il-Yeol Song
Summer 2011

Table of Contents

Abstract

Behavioral targeting (BT), also known as interest advertising, is an online tool utilized by individual web sites or network advertisers to enable tailoring the display of ads to the current online user's interests. Behavioral targeting allows marketers to target the audience who has shown interest in their respective category because this usually guarantees a higher success rate of winning them over as a potential customer. There are two types of behavioral targeting strategies: web browsing behavior and search query behavior. Web browsing behavior keeps track of the pages visited whereas search query tracks the words searched in major search engines (Yan, Liu, Wang, Zhang, Jiang, & Chen,

2009). In this study, Yan et al has shown empirical results supporting the conclusion that using search query behavior outperforms web-browsing behavior as the behavioral targeting strategy of choice. Regardless of the strategy, the nature of the data that drives behavioral targeting is the main interest for this survey.

In 2009, Google has launched its interest-based advertisement. Google users are not aware that their private data are collected each time they are online but the company does provide the option to choose the type of data that is being collected in the background during their online activity. Similarly, Yahoo! followed suit but went to the ultimate of giving users the “opt-out” option. This is due to the many privacy complaints and legal actions taken by consumers against these companies (Helft, 2009). Circa this time, a lot of hype has been generated for this new innovation and behavioral targeting became the latest and greatest tool for target marketing. Google and Yahoo are examples of some of the bigger online presence. Social networking is also jumping on this bandwagon when Facebook adopted behavioral targeting. Facebook has partnered with Blockbuster and eBay in that when a Facebook user is on the site, the activities information is broadcasted to all of the people within the user’s network.

The implementation of behavioral targeting for a company is a big undertaking because it involves collecting a good data sample, defining and understanding the business goal so that engineers can create a predictive model that satisfies the business goal, and investing in technology. This survey will traverse through the various technology aspects of behavior targeting.

This paper covers three main topics: Technology, Data, and Issues. The Technology discusses the emerging tools that enable behavioral targeting. Data focuses on the database technology, models, and issues that exist. Issues section identifies the outburst of privacy concerns radiating among consumers.

Introduction

Zetabytes of data are created each year. As technology permeates the globe, this amount will continue to increase at record rates. Social network sites, retail organizations and government agencies all have a need for this data. Businesses realize the importance of mining said data to individualize marketing efforts and expand the reach into consumers’ needs. Amazon, Facebook and Twitter are examples of organizations that utilize recommender systems and behavioral targeting as precise advertisement tools. We will attempt to show the advantages as well as subsequent disadvantages that result from using behavioral technologies. Additionally, we will describe the technology that is used to achieve successful marketing efforts and list the major companies that employ these technologies. Lastly, we will discuss future impacts on consumer privacy and the levels of apparent quid pro quo that affect common users.

Behavioral Targeting Technology

A survey of the scholarly literature reveals several Yahoo! patents for behavioral targeting systems, state-of-the-art work on social CRM and recommender systems, and a method for achieving efficient and scalable behavioral targeting systems which help businesses accomplish the goal of maximizing advertising impact, reducing advertising budgets, and enhancing customer relationships.

Yahoo! Behavioral Targeting Innovations

On July 15th, 2010, A group of Yahoo! researchers published a patent application for a method and system which addresses "Large-Scale Behavioral Targeting For Advertising Over a Network." This invention is for behavioral targeting over a network such as the Internet, as opposed to behavioral targeting within one particular e-commerce site (recommender systems). Essentially this method and system is a unique application of statistical machine learning.

From a high-level perspective, the method of this invention begins when the system receives training data that has been preprocessed by a data-preparation component which has reduced, aggregated, and merged the data. This training data is essentially raw user behavioral data, e.g., ad views, ad clicks, organic clicks, searches, and/or overture clicks. In terms of size, this data can be on the order of 20-30 terabytes. The goal of the pre-processor device is to reduce the data size at the earliest opportunity without losing important information or repeating it (redundancy). The data-preparation device ultimately reduces the data size to between 2 and 3 terabytes and ensures no loss or redundancy occurs. Using this data, the model training device fits a linear Poisson regression model from the preprocessed data (Chen, 2010). First, the system selects granular events such as page views, ad clicks, and search queries as features (regressers) based on their frequency. Three vocabularies result from the frequency selection process which together define an inverted index of features. This index can then be used to reference a feature name by its index for generation of feature vectors (training examples)- the next step in the overall method. The system uses a data-driven approach to initialize the weights of the targeting model, instead of assigning uniform or random values as many gradient-based algorithms do. Specifically, two separate methods are used by Chen et al. to initialize the weights, namely, feature specific normalization and target-specific normalization. This initialization of model weights is accomplished in a single scan of the training data set. The final step of the overall method is scanning the feature vectors iteratively using multiplicative recurrence in order to update the weights of the BT model. The weight update process is very similar to the initialization method, with some minor variation in the type of normalizer used. The method of this invention involves advanced statistical mathematics, the mechanics of which ultimately determine its effectiveness, efficiency and scalability (Chen, 2010).

Figure A below from this Patent application illustrates the high-level behavior of the invention described briefly above. These behavioral targeting models have been shown experimentally to predict click-through-rate (CTR) from user history by a lift of 20% over conventional systems. Also, prior to the invention of Chen et al's method and system, the conventional standalone modeling system required an entire week to process 60 BT-category models. The technology described in this patent application can create 450 BT models within *one day (*Chen, 2009). This is quite a remarkable improvement in scalability and efficiency. The particular embodiment of this system which was implemented and tested use the Hadoop MapReduce Framework, but other embodiments could be achieved.

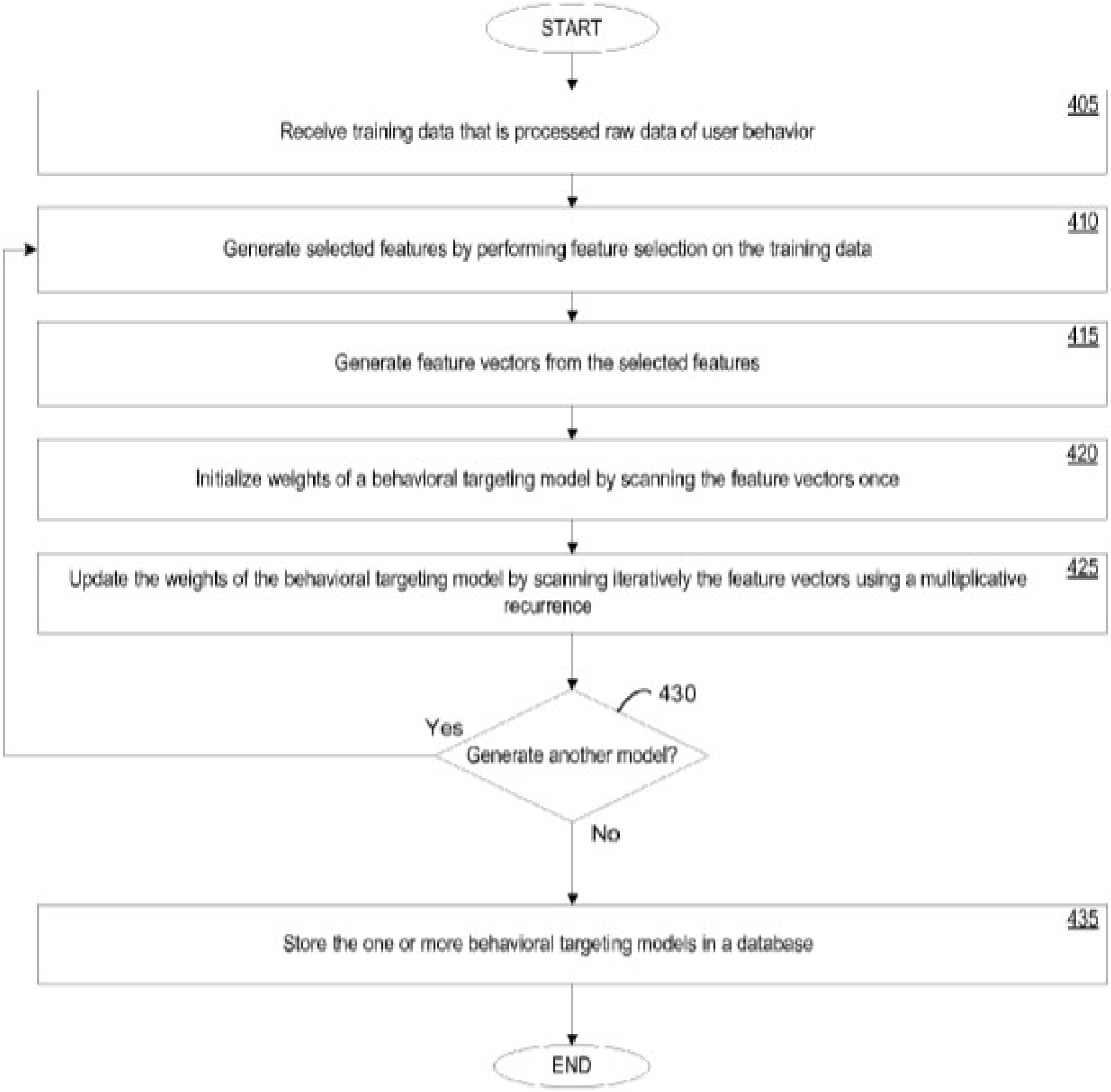

START
Receive training data that is processed raw data of user behavior
405
Generate selected features by performing feature selection on the training data
410
Generate feature vectors from the selected features
415
Initialize weights of a behavioral targeting model by scanning the feature vectors once
420
Update the weights of the behavioral targeting model by scanning iteratively the feature vectors using a multiplicative recurrence
425
430
Yes
Generate another model?
No
Store the one or more behavioral targeting models in a database
435
END

***Figure A*. Flowchart Model depicting Behavioral Targeting Method. Adapted from U.S. Patent Application for "Large Scale Behavioral Targeting for Advertising over a network" by Chen, Y.,Pavlov, D.,Berkhin, P., Canny, J. Google Patents. California. Retrieved from http://bit.ly/qC32Mk. (2010).**

On April 5, 2011, Yahoo! was issued a patent by the United States Patent office for an invention which accepts granular data as input for behavioral targeting using predictive models. This technology is an empirical data mining approach to behavioral targeting which learns from the user's past behavior to predict future actions. It draws different kinds of behavioral data from across the web via information logged in multiple servers within a network of websites. More specifically, this invention is a means of utilizing granular user behavioral data such as ad clicks, ad views, and search queries to generate a predictive model which can automatically learn behavioral patterns of a particular user and possible actions she could perform, and then score and rank the user. For example, this technology can generate a model fone dayor Yahoo! which can score and rank a set of users for the likelihood they will click on a Honda Accord search advertisement based on the granular behavioral data it received from the user's past actions, such as their Yahoo! search engine query for "used cars" and their past click-through of a series of Honda advertisements, e.g.

From a high-level perspective, how this technology works is straightforward. To begin, the method receives multiple data from granular events such as search clicks and ad views, and pre-processes this data to determine the amount of informational content for target prediction. Pre-processing includes (1) pruning the granular event "noise" by eliminating any that occur less than a preset number of times, (2) aggregating the granular events by combining them into a total count for a training period, and (3) clustering the granular events into N groups based on their informational content for target prediction. The pre-processed data is then used to generate a predictive model. The predictive model is used to decide the probability of a user action and includes a weight for this possible action along with model parameters which are linear combinations of the pre-processed data. The method then trains the predictive model by adjusting the weight of a hypothetical user action thereby optimizing the performance of the model. Finally, the method applies the model to selected users to score them; users are also scored by applying a linear Poisson regression model based on the ratio of a predicted number of ad clicks to the estimated number of ad views (Canny, 2011). Figure B below illustrates the high-level work flow of this method.

To overcome the issue of the large-volume of behavioral data, as well as its high dimensionality and sparseness, the conventional approach in behavioral targeting is to generate business-driven categories for granular events; predictive models can then be built on this category-aggregated data. You may recall the previous invention discussed above used such a taxonomic approach in its targeting method. This patent document mentions a few problems with this conventional taxonomic approach, namely, (1) important information resolution can be lost in the categorization process since many different events are often group together in a category, and (2) since the chosen categories are business driven instead of problem-driven, even more information can be lost. With this in mind, Yahoo! decided to bypass using business-driven categories for building their predictive models; instead they build predictive models directly from granular events (Canny, 2011). Therefore, the method of this invention does not rely on the availability of any pre-defined business categories and is more adaptable to the problem domain given such independence.

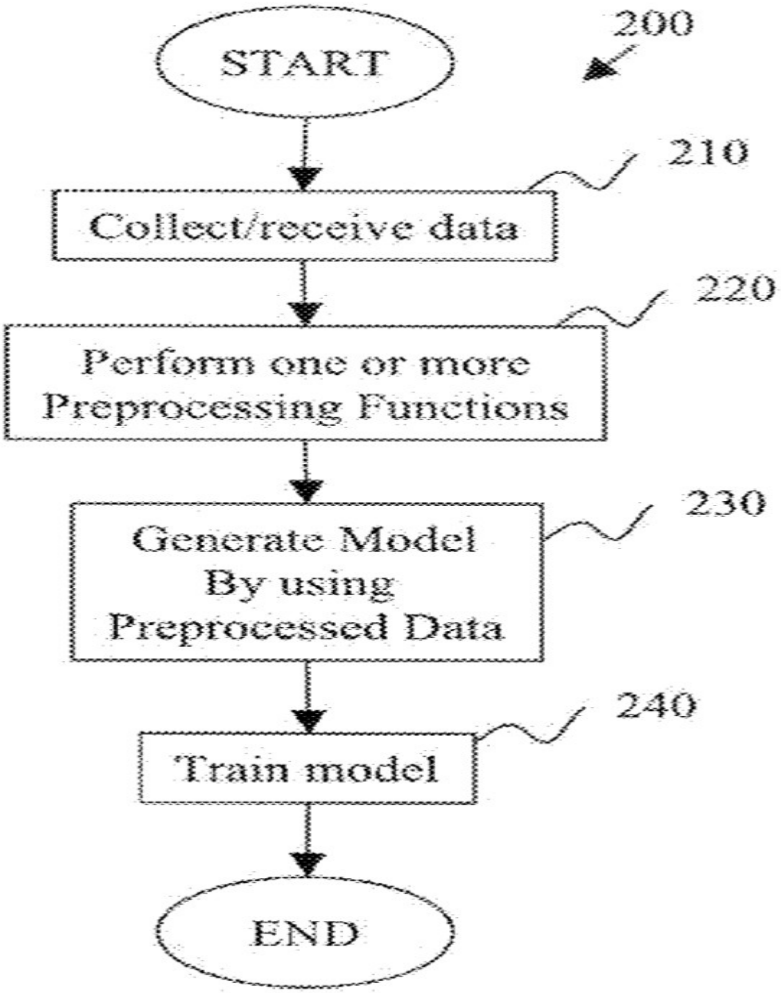


Figure B. **Diagram illustrating Model generation and/or training. Adapted from US Patent No. 7,921,069 Washington, D.C. US Patent & Trademark Office by Canny, J., Zhong, S., Gaffney, S., Brower, C., Berkhin, P, John, G.H. (2011).**

A team of Yahoo! researchers filed another patent with the U.S. Patent & Trademark office on August 25, 2011, for a system which uses a site-sequence value algorithm to deliver the most relevant ads to a variety of visitors to a web portal. This method allows advertisers to overcome the problem that most users who agree to be targeted and served ads are tech-savvy (high engagement level), so their behavior can not be presumed to apply to infrequent users of the Internet (low engagement level).

The site sequence is defined as a recorded sequence of events which occur between site A and site B for at least one person with a corresponding engagement level. This technology studies visits from the first site to the second from a variety of users and measures the average engagement level. The method also studies visits from site B to site C by measuring the amount of time the user remains on site B so that it can determine level of interest. Another important aspect of this method is that it determines the likelihood a user will be interested in ads displayed on site C (Gerster, 2010). Figure C below illustrates the high-level site-sequence method.

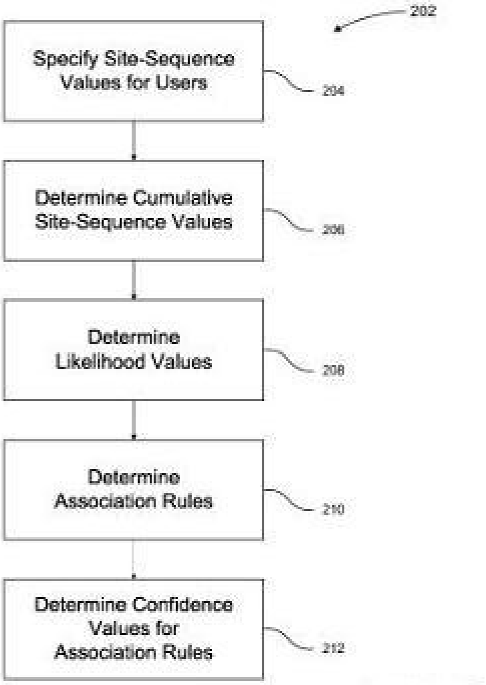

202
Specify Site-Sequence Values for Users
204
Determine Cumulative Site-Sequence Values
206
Determine Likelihood Values
208
Determine Association Rules
210
Determine Confidence Values for Association Rules
212

***Figure C.* Diagram illustrating site-sequence value algorithm. Adapted from U.S. Patent Application for "Forecasting Association Rules Across Engagement Levels." by Gerster, D., Awadaliah, A, Thampy, S. Google Patents. California. Retrieved from http://bit.ly/rfc6V5 (2010).**

Commercial Status of the Yahoo! Innovations

While only one of these technologies has been officially patented, they have all been implemented into the Behavioral Targeting product as part of Yahoo! display ads product portfolio. Yahoo! does not license it to other companies for competitive advantage reasons (C.Y. Chung, personal communication, August 17, 2011). Yahoo! offers a unique behavioral targeting product that addresses the needs of both Brand and Direct Response advertisers. Using the technology described above, Yahoo! provides advertisers with both "upper-funnel" audiences who are beginning product research and "lower-funnel" audiences who are near to making a purchase (Yahoo!, 2011).

Challenging Research Issues Related to Yahoo! Innovations

A major challenge of current research at Yahoo! is that user activity with respect to advertisements is relatively sparse. Users rarely click on ads, and when they do, purchase is even rarer. Given that the click-through rate is less than one-tenth of one percent, researchers have to collect extremely large amounts of data in order to predict interest precisely (C.Y. Chung, personal communication, August 18, 2011). Finding effective methods to collect, model, and process this sparse data to predict user's interest in advertisements is an ongoing challenge for researchers from Yahoo! and other major online companies.

Inferring Advertisement Receptiveness

In 2009, Guo et al. described a method for inferring a user's receptiveness to search advertising using models for search context and behavior; test results have shown this method significantly improves the ad click-through prediction for the user compared to methods which rely on classification alone without modeling this additional information. This method addresses the timing of ads, an aspect of search advertising not previously covered in practice or the literature. The limitation of previous methods is that without timing the ads properly, systems can inadvertently annoy the user and 'train' them to ignore ads; conversely, systems may not serve ads at times when users are, in fact, interested in viewing them. This technology actually predicts whether a user would find interest in advertising for future searches within the current session, before the user issues the queries. Given these predictions, the search engine can take them into account when deciding what ads to display and how many of them. This promising technology could have a positive impact on advertising budgets since ads would only be displayed when the user is, "in the mood" for them.

Rodden et al. showed that mouse movements correlate well to the user's eye movements; Guo et. al.'s model uses mouse movements as an inference about users short-term interest and attention. In this system, the mouse-movement data is collected using Java Script code embedded into headers of the search engine result page. In the session-level contextual model, two hidden states represent ad-receptiveness (receptive or non-receptive), and using a sequence of related searches as input, the model infers the most likely hidden state and predicts ad-click through based on the last inferred state of the sequence of related searches.

Guo et al developed a prototype named the Contextualized Interaction modeling system (CxI) which captures context features such as search queries, and interaction features such as mouse activity, scrolling activity, and menu and button activity, in addition to page dwell time, and click through on results. The most important feature of this model is the various representations for mouse trajectories as follows:

- Different ranges of trajectories are considered in the model as global features.
- The trajectory features model precise physiological characteristics such as speed, acceleration, and rotation of the user's mouse movements.
- A Boolean hover feature which is true if the user hovers over the north or east ad-display region of the search-result page for 500ms.
- Ad click through on the current page is also measured as a binary feature in this model.

Session level classification of the user's receptivity to ads (i.e. receptive or non-receptive) is defined based on the conditional probability of these states given a set of observed sequences of searches (Guo, 2009).

This technology is still in the research phase and has not been pushed into commercial production yet. The most challenging research problems related to this technology are how to formalize the problem and how to model the search interaction and context (Q. Guo, personal communication, August 21, 2007).

Social CRM and Recommender Systems

Within the social networking world, Customer Relationship Management (CRM) is best accomplished by efforts to improve the user experience within these networks. The logic is that if a customer enjoys their experience using this social network, then they are likely to remain a customer, that is, spend more time on this social network instead of others. Consequently, an important business and technical challenge for social networks is making the user experience as interesting, fun, and 'addictive' as possible.

The key feature to improve in social networks such as Twitter, Facebook, and Google+ is the conversation stream because conversation has been empirically shown as the primary reason for using social streams (Java, 2007), and finding interesting conversations without machine intervention can often be difficult due to sheer information overload. The main problem with conversation streams is that, for example, a Facebook user may encounter over 150 conversations in their news feed every day, most of which they have neither the time nor interest to read (Chen, 2011). Hence, developing systems which serve the most interesting conversations to users and filter out noise is of tremendous importance.

Since users of social streams have different purposes for using these networks, namely, social or informational purposes, there is a diversity in conversation preference. For example, faced with the choice of reading either Jim's conversation about climbing mount Everest or Wendy's recent two-week vacation in Germany, Alex may prefer the latter because he cares more about Wendy while George may prefer the former because he is an avid mountain climber. This diversity in conversation preference poses a technical challenge to developing recommender systems since the algorithms which drive the system require different assumptions to serve the most relevant content.

Chen, et. al. studied the application of six different algorithms using some combination of three distinct conversation factors to Twitter's conversation stream and documented the outcome by surveying a representative set of users.

How the system works is rather interesting. First the system gathers all conversations a user's followees have participated in, including conversations between followees and non-followees, as candidate conversations. Then the system ranks conversations using algorithms which compute based on thread length, topic relevance, and tie-strength (closeness of association between user and followee).

The thread-length factor is determined by the number of tweets in the conversation. Topic relevance, on the other hand, is determined by first building a topic profile vector for the user- which is essentially a set of weighted words the user has used in tweets. The next step in determining topic relevance is to build a similar vector for the conversation content. As a final step in determining topic relevance, the system matches these two vectors and calculates a relevance score. Tie strength is determined separately by the occurrence of direct communication between users, the frequency of direct communication, and the tie strength between two users and their mutual friends.

The first algorithm (*Random*) used in Chen et al's study recommends random conversations and is used as a baseline for other algorithms. The second algorithm (*Length*) ranks conversations based on their thread length, ie, conversations with highest thread length given priority. The third algorithm (*Topic*) ranks conversations based on their topic relevance score, which is computed based on the procedure described above. The fourth algorithm (*Tie*) recommends conversations based on their *average* tie-strength score; for example if John has a stronger association with Jenny than Bill, then Jenny's conversation will be recommended before Bill's. The fifth algorithm (*Tie-Sum*) recommends conversations with the highest *total* tie-strength scores. In this approach, the algorithm ranks conversations by computing the product of thread-length and tie-strength (a hybrid of the *Length* and *Tie* algorithms). The sixth and final algorithm (*Topic-Tie-Sum*) recommends conversations with the highest product of the topic relevance score from *Topic* and the total tie strength score from *Tie-Sum*. This final algorithm can be viewed as a combination of all three factors described above: thread-length, topic relevance, and tie-strength(Chen, 2011).

After running tests with a group of 35 Twitter users and the six algorithms described above , Chen et al found that the five main algorithms performed better overall than the random baseline algorithm, and the sixth algorithm which combines thread-length, topic relevance, and tie-strength, was perceived by the test subjects to have recommended the most interesting conversations overall. These algorithms do not take into account the diversity of usage purpose described above, so creating systems which somehow infer the purpose for using the social network- social or informational- is the basis for future work in this area.

The lead researcher Jillin Chen explained in a personal email that the Palo Alto Research Center owns this intellectual property and will not disclose the commercial status of the research because it is considered a trade secret; therefore, the commercial status of this technology is currently unknown. The major players involved with research of this kind are Microsoft Research, Google, Facebook, and Twitter, and they tend to keep the commercial status of their trade secrets fairly tight lipped (J. Chen, personal communication, August 17, 2011).

Big Data

The advent of the World Wide Web not only sparked the oncoming of the information age for individuals but it has evolved to become a critical tool in the twenty-first century for businesses of all sizes. Data is now the focal point and driving force for both the technology and marketing industries. The easy availability of consumer's private data, including their interests, preferences, and online activity becomes a digital mine that contains a wealth of information for businesses to use to their advantage. This data collected is called clickstream data and it constitutes what behavioral targeting centers on.

The size of clickstream data is massive in size through the enormous accumulation of data collected from different ubiquitous channels. Since the web is accessible almost everywhere by mobile phones, tablets, laptops, and traditional computers, this growth is daily. With today's storage technology, that physical storage is no longer a big issue but the manipulation and interpretation of the data is what challenges technologists and business analysis today. In addition, Web-scale data is changing so rapidly that keeping up with the newest, most recent data requires new capabilities to be repeatable and fast performing to keep up with its growth. This type of data is dubbed as “Big Data”.

Clickstream big data are valuable assets for advertising agencies and independent online websites in understanding each web visitor's interests and tailoring the display of ads to fit their interest categories based on historical navigation and keyword searches.

The remainder of this survey will mainly focus on revealing the details of various technologies and tools used to massage, analyze, interpret collected user data. It will also discuss the major issues and challenges incurred with big data. The intention of this discussion is to present an overview of the popular and common data technologies involved in behavioral targeting or other goals similar to it.

We will continue the survey with the following topics:

- Data infrastructure that supports behavioral targeting. This subtopic will discuss a holistic view of the infrastructure, including how data is inputted and outputted for BT.
- Data technologies/tools. The heart of understanding and utilizing big data lies within data mining. This section will primarily focus on data mining and processing of the data.
- Data Storage. Due to the need for faster retrieval and turn around time for data analysis, database systems for storing such big data are not conforming to the typical relationship database management systems (RDBMS).
- Current Challenges/Future Advancements. Although data technology has improved tremendously since the inception of databases, current challenges exist with managing data anomalies such as incomplete data, missing cookies, et cetera.

Infrastructure

Web-scale data, by nature, requires the supporting infrastructure to scale to a large amount of data and most importantly, fast performance. Infrastructure includes the appropriate hardware and software for a technology goal or implementation. In terms of hardware, processing such data utilizes parallel computing in which each machine in a single or multiple cluster(s) takes part in working on a task (a divide and conquer approach). Software includes data mining approaches, algorithms, and data analysis tools that will drive the process.

In “Practical Lessons of Data Mining at Yahoo!” Yahoo Inc. reveals some of their lessons learned over the years. Yahoo! is one of the leaders today in pioneering the customization of displaying ads to its online users. The authors describe the hardware requirements for Yahoo’s BT application and recommend them as the basic necessities. The list consists of these basics: machine clusters running a parallel framework such as Apache Hadoop, a distributed file system, and a lot of disk space to store and process multiple sets of data (Chen, Pavlov, Berkhin, Seetharaman, & Meltzer, 2009). Yahoo's infrastructure represents a generic architecture at the macro level so each implementation for big data processing can vary. The general guideline to implementing parallelism is to distribute the tasks very finely so that better randomization would reduce the likelihood of a bottleneck in the cluster nodes.

To better understand this notion, we will explore Google's infrastructure which not only supports interest marketing but also its sixty plus products. Google acquired DoubleClick (an online advertising firm) a few years ago and has implemented its own version of interest-based advertising. Web users on Google's suite of applications will have the option to select preferences for their online navigation data that is automatically collected. The caveat is there is no way of opting out of having data collected ("Google Begins Behavioral Targeting Ad Program"). The total collected data changes frequently as new browser cookies are captured, along with tracking all the deltas by time.

Similar to Yahoo, Google also runs on MapReduce. Its main infrastructure setup is a master server that delegates and distributes work to its tablet servers. The tablet servers are added or removed dynamically to a cluster to adjust the workload. The MapReduce program is used to efficiently generate key/value pair mappings for the input data and reduce the pairings' set of values into a single value so that data can be retrieved at a faster rate. Google's infrastructure centers highly on its own proprietary in-house developed applications. The storage of web-scale data will be discussed in details in the next section.

Overall, the general infrastructure that supports web-scale data relies heavily on parallel computing and frameworks. Parallel frameworks refer to a class of data center application support tools that can be used to support highly parallel data analysis. The programmer will supply the kernel of the computation and the framework automatically invokes many copies of the kernels in parallel and controls the overall execution. The copies will mostly like run in parallel in multiple clusters (Gannon).

Data Technologies

1. Data Storage

The storage of web data, especially clickstream, is innovative and contrary to traditional database systems. Before discussing the storage methods, we want to first explain the format of clickstream data is and where the data comes from. The exact tracking of an online user's navigation within a site's domain comes from the web server log file. A file is produced each time a web site is accessed (Alves & Belo). Clickstream data contains the following elements:

- IP address of the user's origin site

- Access time
- Referring site
- URL of the target site
- Broswer method
- Protocol that was used

These elements becomes the source for companies to figure out usage pattern on the website and most importantly, to understand what the users' behaviors are. The process of extracting knowledge from this data is known as Web Mining and tools that perform such analysis are available as freeware and commercial clickstream analysis tool. More details on the various technologies used to mine clickstream data is discussed in later sections.

a. Bigtable for Big Data

When the word "speed" is mentioned within the context of the web, Google's name is bound to surface. In terms of fast searches, the general public knows the company for its algorithm. However, what is not yet popularized is its storage system. Google is a multi-platform multi-application suite available on the Internet. With the recent acquisition of DoubleClick, Google now provides digital advertising such as ad serving services. Its clients include Nike, Visa, Microsoft, Coca-Cola, Apple Inc., and more ("Doubleclick").

Google stores its Big Data in its proprietary distributed storage system called Bigtable. It is a special database management system built in-house as a solution to house its large, unique data. In a Google publication by Chang et al, the authors use Personalized Search as an example to briefly explain how Bigtable works. Personalized search is another Google product that collects user data while the user is on a Google site. The user profiles are stored in Bigtable and are indexed by a unique userid by a row. The user actions are stored in family columns and the timestamp stores the time an action occurred. When a client calls the Bigtable API and queries for some data, Bigtable returns a flexible data model. The calling application can dynamically control the data layout and format of the returned data. This data model, which is the storage and retrieval format, is known as SSTable. Before diving into a discussion on SSTable, we first discuss the general architecture of Bigtable. The system comprises of mainly the Google File System and SSTables. MapReduce usually intercepts the data and processes it prior to calling the Bigtable API. All of these processes are managed by its own cluster scheduler, called Chubby. The diagram presented below will put the aforementioned into context and give a holistic view of the components discussed.

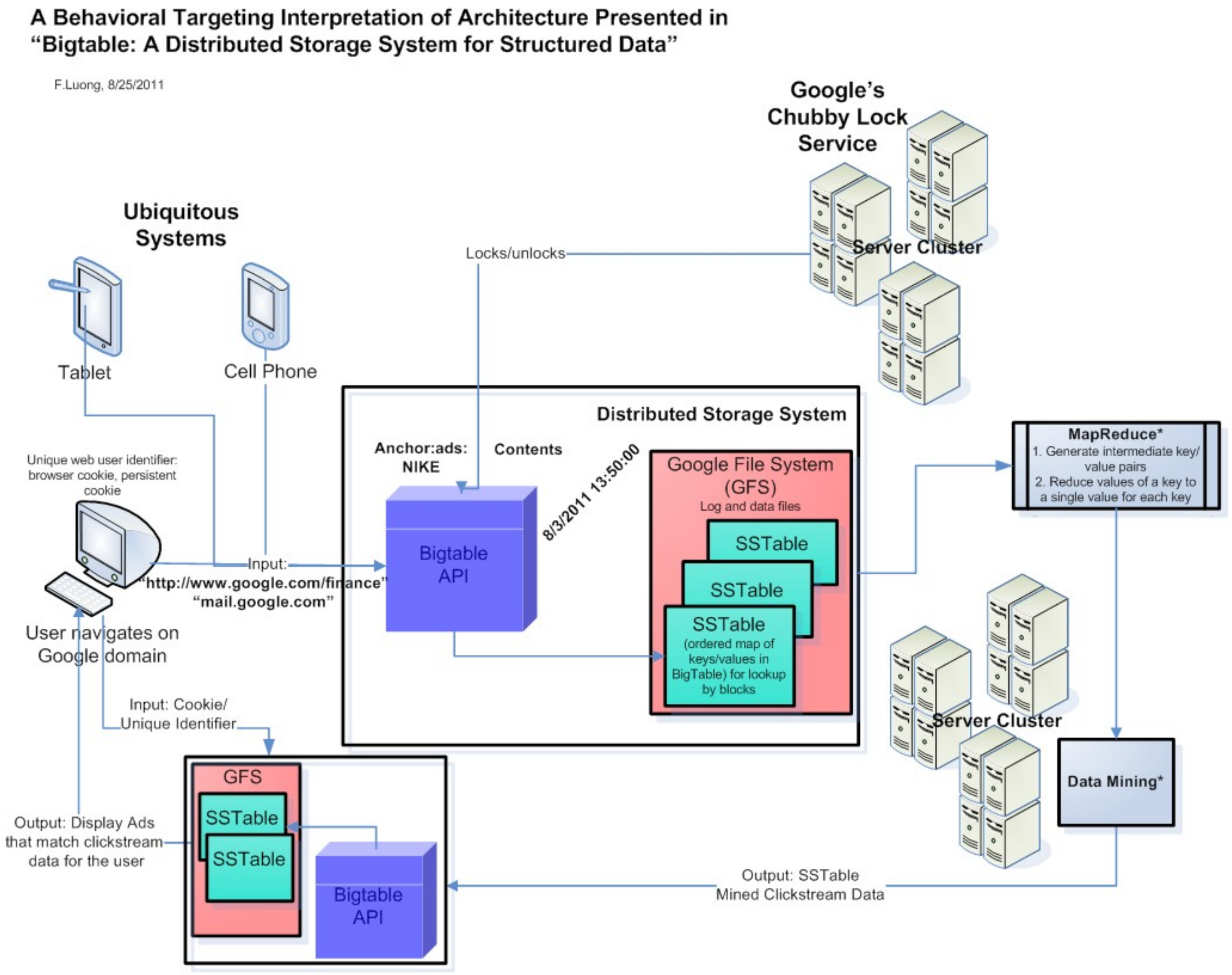


***Figure D1*- Interpretation of how Bigtable, MapReduce, and Data Mining works together for behavioral targeting implementation. This is a conjecture based on readings for this survey.Google has not disclosed technical details on how it implemented behavioral targeting (BT). From the content presented in the paper, the data structure on storing big data is revealed along with its MapReduce process. To put a behavioral targeting perspective on this architecture, an interpretation is conjectured based on the facts in the paper and common BT notions that were collected in the research of this survey.**

The diagram shows that clusters of servers are heavily utilized in the infrastructure. This is a point that resonates throughout the paper and is further confirmed in other white papers about Google. The starting point of the diagram is the ubiquitous systems. Users can access the Google domain (applications within google.com) from a computer, portable tablets, or even cell phone these days. The users are tracked given a unique identifier in the form of a browser cookie. The user's navigation on the website and the content of the site are captured and stored via the Bigtable API. The main data model for storing multifaceted and related big data is SSTable, a file format that stores the data in key/value pairs. The format embraces the Row:Column:Time structure.. A Bigtable is represented by one or more SSTables.

For the purpose of BT, the clickstream data of interest is the clicking of the ads on Google applications such as Finance and Gmail. A process called MapReduce will take all inputs from different sources, create keys and intermediate values for the inputs, and then aggregate all the values for each key, which becomes the final key and value. Conceptually, the MapReduce can be run on clickstream data, defining origin website and ad clicks. The reduced and final data can then be mined with various mining techniques to predict user interests and match the appropriate ads to display for the unique identifier for a user. The MapReduce and data mining technologies are discussed in the latter parts of the survey.

Bigtable, when viewed as a composition of multiple components, is considered as a distributed storage system but as a storage entity alone, Bigtable is a "sparse, distributed, persistent multidimensional sorted map" (Chang, Dean, Ghemawat, Hsieh, Wallach, et al., 2006). On a conceptual level, Bigtable is an API that is exposed to inserting and retrieving data from SSTable. Each value in the map is an uninterpreted array of bytes. This means that the system (machine) does not care about the data and does not need to know what the data means. Occupancy is what matters. A map's content is indexed by 3 dimensions: [row: string], [column: string], and [time:int64] as shown in the picture below.

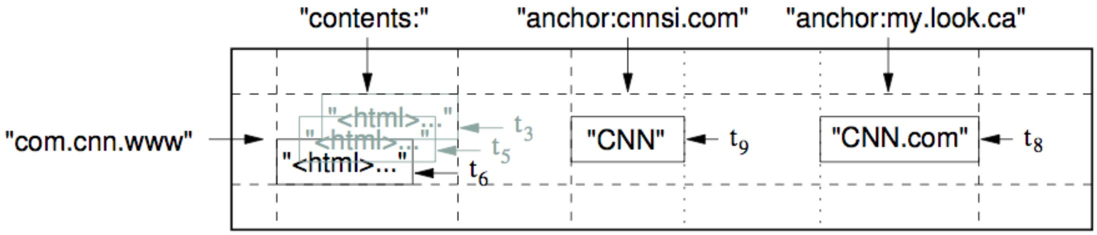


Figure D2- **A slice of an example table that stores Web pages. The row name is a reversed URL. The contents column family con- tains the page contents, and the anchor column family contains the text of any anchors that reference the page. CNN's home page is referenced by both the Sports Illustrated and the MY-look home pages, so the row contains columns named anchor:cnnsi.com and**

anchor:my.look.ca. Each anchor cell has one version; the contents column has three versions, at timestamps t3, t5, and t6.

Reprinted from "Bigtable: A Distributed Storage System for Structured Data" by Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber

Storage Model for Bigtable		
Row	**Column**	**Time**
• String type of max size 64 KB • Read/Write to a single row key is atomic so either all or nothing of the row will get updated regardless of how many columns are read or updated. • Row key is ordered by lexicographical order (alphabetical order). • Rows are partitioned into separate tablets for each row range. • Row keys are usually URLs.	• Columns are grouped by column families. • Columns in same column family store the same data types. • Column family must first be created before any data can be stored in any column. • Column key is determined by family:qualifier. i.e. language:English.	• Main differentiator for same set of row-column data. Bigtable can store duplicate data but their instances can be determined by the timestamp. • Versions of data are stored in decreasing timestamp so the most recent is stored first. • Data management can be done automatically by Bigtable or by the client.

***Table D1*- The table lists the attributes for each dimension of the storage model for Bigtable.**

SSTable is a file format that Bigtable uses to store and look up data. It is "an ordered immutable map from keys to values, where both keys and values are arbitrary byte strings" (Chang, Dean, Ghemawat, Hsieh, Wallach, et al., 2006). An SSTable contains a sequence of blocks of size 64 KB and the block index is stored at the end of the file to locate blocks. Operations are exploited to search for a value that is associated with a specific key or to traverse through the key/value pairs within a key range.

Bigtable's multidimensional design is flexible in regards to storing data and multiple copies of the same data. This survey will not further discuss Bigtable but a reading of the published paper is recommended for those who are interested in the details around Bigtable's cluster nature and additional functional requirements such as caching, transactions commitment, performance, et cetera.

b. Amazon's S3

Amazon introduces a new revolution in data storage to the technology community. It offers its clients distributed data storage for a service fee. Simple Storage Service (S3) is Amazon's proprietary storage mechanism offered to the public but Amazon also stores its own web-scale data in S3. Therefore, it is worthwhile to explore this type of data storage because of its unconventional availability. It enables small businesses that do not have the capital to invest in technology to perform data analysis, mining, and customer targeting.
The structure of S3 storage is rather simple. It is organized over two levels of namespaces but the main storage model is a bucket object. A bucket is named appropriately for what its purpose is because it acts like a bucket that stores anything that is given to it. A bucket is similar to a folder or container concept. It stores an unlimited number of data objects. The bucket and data object are data structures specific and available only to Amazon.

Bucket: unique name, data object(s)
Data Object: metadata (user-specified key/value pairs and predefined HTTP metadata up to 2KB), blob (up to 5 GB)

The advantage to this type of storage is that the client can store data in any form since the nature and type of data required by each company can vary vastly from one another. However, the drawback is maintaining existing data objects will require that the object be downloaded, modified, and then reloaded back to S3 (Palankar, Onibokun, Iamnitchi, & Ripeanu). Search capabilities are also limited. Aside from its shortcomings, S3 is still a useful mechanism for data-intensive system for storage and further system processing.

The availability of S3 is unconventional because the entire platform is on the cloud, whether it is a private or public cloud. Unlike Bigtable, which is proprietary to Google, S3 is accessible to the public. For a monthly service fee, clients have access to an environment that has the hardware and software set up for data uploading and downloading. The client will only have to use the Amazon provided interface (whether it is ETL or another technology) and not have to bother with investing in hardware, software, IT professionals, and support.

Although the storage model of S3 is attractive, one primary concern for this big data storage is security. Unlike Bigtable or a typical relational database, the security of S3 is governed by Amazon and not within the security mechanisms of the company using it. Further evaluations of security are discussed in the article "Amazon S3 for Science Grids: a Viable Solution". This paper uses an application that deals with scientific big data to experiment with S3 because the science community is churning terabytes of data each day.

In relation to behavioral targeting, companies can store its clickstream data in S3 in the desired format and then use S3 as the source for extracting knowledge and perform ad-user matching. Another option is to utilize Amazon's MapReduce service. Amazon realizes the need for companies to not only connect with their customers but to also have a deeper understanding of what their customers are interested so they can tailor their business activities or strategize their profit goals accordingly. In order to provide a full suite of business services, Amazon offers their Elastic MapReduce service as part of the Cloud suite. It uses the Hadoop framework as the main MapReduce engine and is integrated with processing data stored in S3.

In summary, the storage model for clickstream and big data relies on innovative and new database systems other than RDBMS. The S3 data model and Google's SSTable are both generic enough to store and retrieve data in most industry domains. The webtable presented in the previous sections is only an example used for web navigation data but Bigtable's usage is not limited to web-related data. These two data storage model are two of the popular ones currently. Variations of Bigtable or its fundamental concepts have been implemented in other open-source projects.

We will discuss the topic of data mining next. Data mining is a crucial component in behavioral targeting. If storage is the input to behavioral targeting then data mining is the churning engine that outputs customized display of ads for each individual online user.

2. Mining for Knowledge

Data mining is not a new concept nor is it new technology. It has been around for more than a decade, dating back to the year 2000. Old data mining revolves around mining retrospective data, meaning the knowledge derived from the mining gives a summary of what has occurred (Rygielski, Wang, & Yen, 2002). However, since the year 2000, data mining is aiming to answer business questions that look for predictions and future outlooks. The new mining paradigm is more relevant today because marketing and business is being pioneered to a whole new level due to the ever-changing innovations of mobile technologies and the Internet. The spectrum of advertisements that originally consist of advertisements to the mass through media now includes reaching out to customers on a 1-on-1 basis. This notion forms the foundation of interest marketing in social media, customer relationship management, and online websites. Social media is definitely a trending topic with the boom of Facebook, Twitter, and now the aspiring Google+. However, this survey will focus on customer relationship management (CRM) because behavioral targeting resides in the first two phases of the CRM cycle (explained in sections below).

The second half of this survey is organized into the following topics of discussion: overview of clickstream and CRM, the technologies including data mining models and techniques, data analysis tools, and the standard mining methodologies. We will end the topic on big data with details on issues, challenges, and future advancements.

a. Clickstream and CRM

Clickstream, alluded to earlier, is the sequence of clicks collected from web sites and is pivotal in analyzing a particular user's navigation behavior. It logs information on the particular ads the user has clicked on and other pages on the website that were visited. This type of analysis requires both human and machine analysis. Machines are so powerful today and play such a huge role in analysis that we forget that with extracting knowledge from data, there should be people who will make sense of the extracted information, verify against or demise business hypotheses. The goal of collecting this type of data is to support behavioral targeting. Customer relationship management systems, which fit into the last two phases of the CRM cycle, focus more on the concept of retaining customers and not aiming to acquire new customers. Behavioral targeting aims to make a one-on-one contact and hoping for the user to be attracted to one of the ad's company based on what he/she is interested in, of which the ultimate goal is to acquire new customers.

b. Enabling Mining Technologies

Data mining is a wide field in data technology in which there are many specializations that enables further data processing. Data mining is a key component of a larger process called Knowledge Discovery in Databases. We will not discuss the entire process but focus on data mining because it is the step that extracts useful knowledge that is used to answer business questions. In terms of broadness, data mining can be applied to data in almost any industry. Its main components are mining algorithms (techniques) and mining functions (data mining models).

The CRM cycle follows a framework, which consists of 4 phases: customer identification, customer attraction, customer retention, and customer development. We will discuss only commonly used techniques and functions for the customer attraction and customer identification dimensions of customer relationship management because the latter two phases do not coincide with the goals of behavioral targeting (BT).

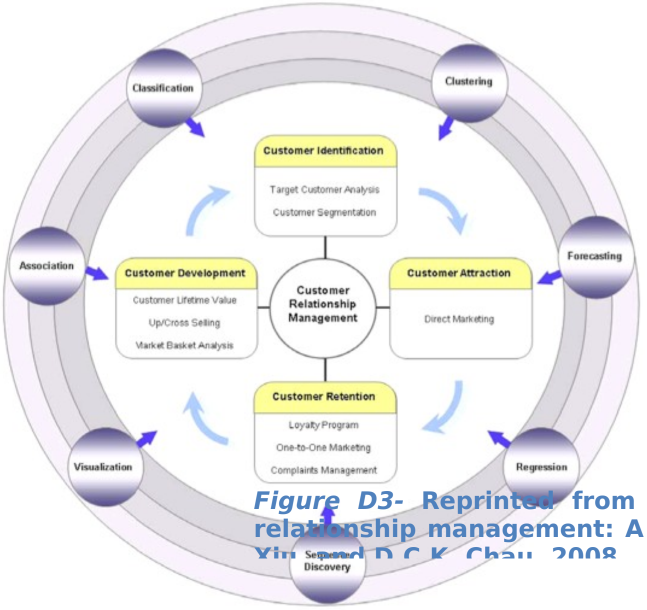


Fig. 1. Classification framework for data mining techniques in CRM.

Figure D3- **Reprinted from “Application of data mining techniques in customer relationship management: A literature review and classification" by E.W.T. Ngai, Li Xiu and D.C.K. Chau, 2008**

Behavioral targeting’s goals align with the first two phases of CRM, Customer Identification and Customer Attraction.

The identification phase involves analyzing and targeting the population who are most likely to become customers or most profitable to the company.

The attraction phase involves the efforts put into attracting the identified targeted customer segments in the previous phase. Efforts can include direct marketing or behavioral targeting.

According to a scholarly survey of data mining techniques completed in 2009, Ngai et al identified the data mining approaches that have been applied to systems that align with the four dimensions of the CRM cycle. They retrieved articles from various online journal databases and reviewed each article to ensure that data mining techniques are indeed applied to CRM. From the reviewed articles, the authors provide a summary of the data mining techniques used by each CRM dimension and report on the top common ones. The table below shows a summary of the authors’ findings but only identification and attraction dimensions are shown in the tables. We chose to focus on these two dimensions because BT’s goal is to identify the interested party (browser cookie identity) and try to win them over as a customer (via customizing display ads based on the party’s previous clickstream). As summarized in figures 2 and 3, classification and clustering are the top two data mining model applied and among the models, neural network, decision tree, and association rules are the most commonly used mining algorithms. A detailed discussion on the data mining process is presented in the methodologies section, which will put into perspective how the model and algorithms fit together.

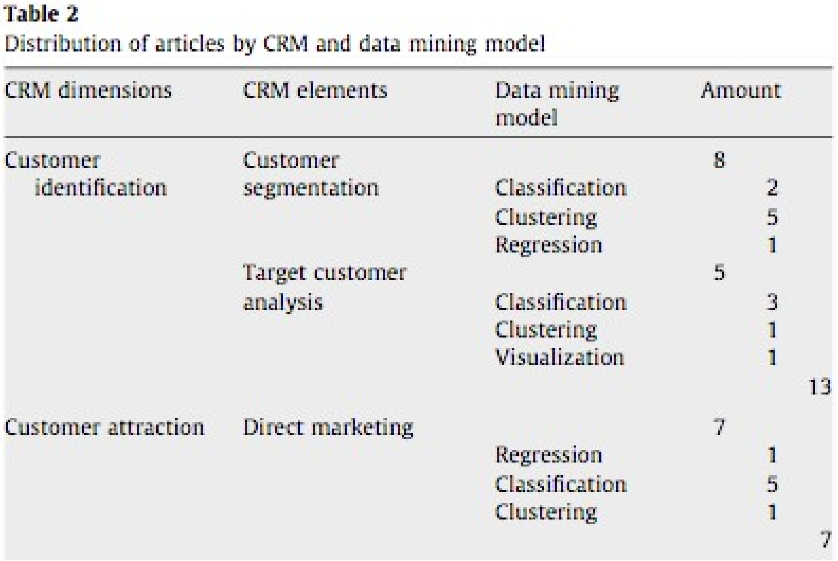


Table 2

Distribution of articles by CRM and data mining model

CRM dimensions	CRM elements	Data mining model	Amount
Customer identification	Customer segmentation		8
		Classification	2
		Clustering	5
		Regression	1
	Target customer analysis		5
		Classification	3
		Clustering	1
		Visualization	1
			13
Customer attraction	Direct marketing		7
		Regression	1
		Classification	5
		Clustering	1
			7

***Figure D4-* Table 2 shows that the two most commonly used models are classification and clustering for both customer identification and customer attraction.**

Reprinted from "Application of data mining techniques in customer relationship management: A literature review and classification" by E.W.T Ngai, Li Xiu, D.C.K. Chau. 2008.

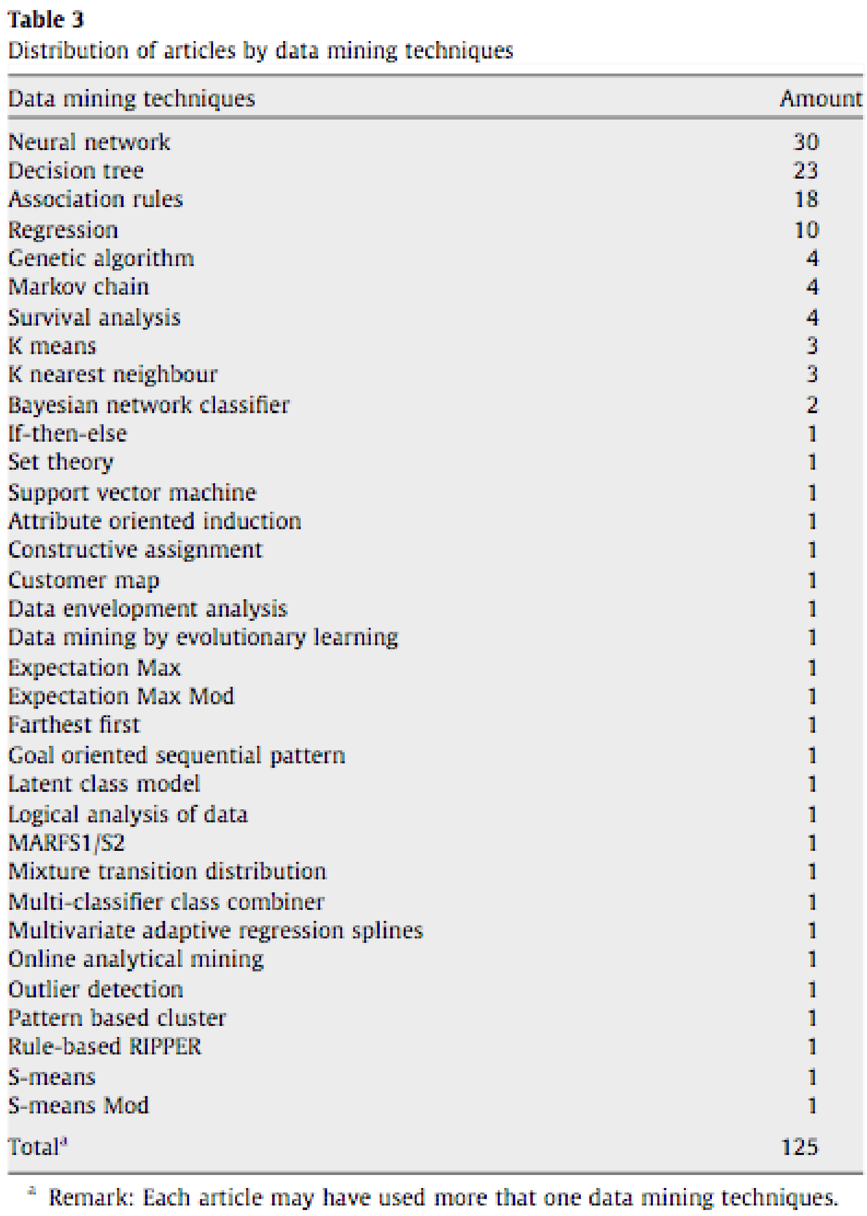


Table 3

Distribution of articles by data mining techniques

Data mining techniques	Amount
Neural network	30
Decision tree	23
Association rules	18
Regression	10
Genetic algorithm	4
Markov chain	4
Survival analysis	4
K means	3
K nearest neighbour	3
Bayesian network classifier	2
If-then-else	1
Set theory	1
Support vector machine	1
Attribute oriented induction	1
Constructive assignment	1
Customer map	1
Data envelopment analysis	1
Data mining by evolutionary learning	1
Expectation Max	1
Expectation Max Mod	1
Farthest first	1
Goal oriented sequential pattern	1
Latent class model	1
Logical analysis of data	1
MARFS1/S2	1
Mixture transition distribution	1
Multi-classifier class combiner	1
Multivariate adaptive regression splines	1
Online analytical mining	1
Outlier detection	1
Pattern based cluster	1
Rule-based RIPPER	1
S-means	1
S-means Mod	1
Total[a]	125

[a] Remark: Each article may have used more that one data mining techniques.

Figure D5- **Table 3 shows that the top 3 mining algorithms are neural network, decision tree, and association rules.**

Reprinted from "Application of data mining techniques in customer relationship management: A literature review and classification" by E.W.T Ngai, Li Xiu, D.C.K. Chau. 2008.

In data mining, a developed model is important in determining how the data should be analyzed. Models satisfy the goal of prediction, the event of using some variables or fields in the database to predict unknown or future values of other variables of interest (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The specific configurations of the model and building of the model that includes rules and parameters can vary depending on the data-mining engine or tool used. It is important to keep in mind that the model's goal is what drives the design and configurations of the model. Classification and clustering both have some similarities. Classification aims at building a model to predict future customer behaviors through classifying database records into a number of predefined classes based on certain criteria while clustering aims to segment a heterogeneous population into a number of heterogeneous clusters (Ngai et al., 2008). Unlike classification, clustering does not have predefined clusters prior to running the algorithm. The clusters are determined from the data and can be mutually exclusive or overlapping. Both of these techniques are useful in mining clickstream data or keywords searched in search engines by web visitors. The decision to use clustering or classification depends on what type of behavioral targeting is used. Clustering is more relevant for network advertisers because advertisers are interested in knowing the entire clickstream trend of all the users for their clients in which predefined classes are not too critical. On the other hand, an individual website is interested in particular classes that are predefined by the business and to answer particular business questions.

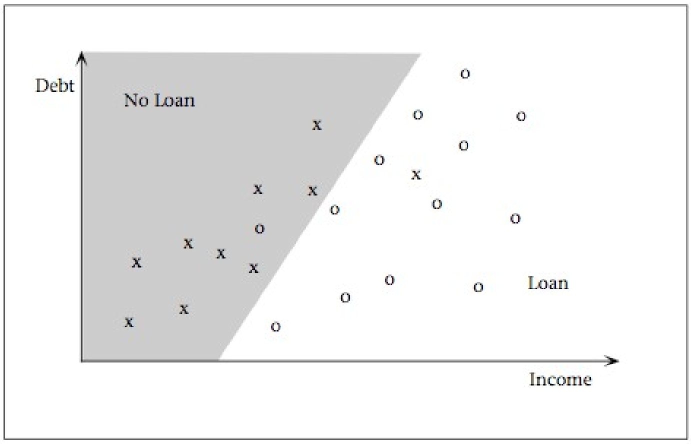


Figure 3. A Simple Linear Classification Boundary for the Loan Data Set.
The shaped region denotes class *no loan.*

Figure D6- **Loan and No Loan are predefined represented by X and O.**

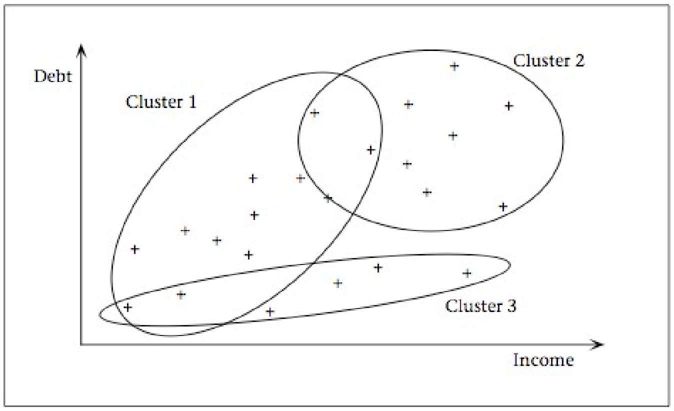


Figure 5. A Simple Clustering of the Loan Data Set into Three Clusters.
Note that original labels are replaced by a +.

***Figure D7-* Clusters are created based on the nature of how the data fits without any predefined clusters.**

Reprinted from "From Data Mining to knowledge Discovery in Databases" by Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996.

The clustering and classification are predictive models that primarily use decision trees and association rules. These algorithms produce simpler representation of the data but their restrictions to a rule or a tree can reduce the types of classification boundaries that can be induced (Fayyad, 1996). Decision trees are structured like a typical tree structure with a root as the topmost node and internal nodes that pose a question. The leaves branching off of the nodes are all possible answers to that question. Decision trees are not binary trees but each node should be a question that is asked to produce the most distinct number of classes. The more popular decision trees for data mining are ID.3 and C4.5. We will not discuss the algorithms in detail but we encourage further readings in sources found in Ngai et al. Association rules are used to test how likely the rule will occur again or how strong the pattern exists within the dataset. This algorithm aims to discover association relationships beyond the obvious among the data. These two algorithms are more commonly used in data-mining for non-experts particularly business people.

Neural networks are more advanced because it can universally represent a function to any degree of accuracy but not without the additional complexity of variables injected into the algorithm. Neural networks can handle categorical and continuous independent variables and it works well with lower cardinality variables. It can be used to solve sequence prediction problems and estimation problems (Rygielski et al., 2002).

Mining BT data generally involves analyzing web log data and browser cookies that consists of IP address, URL of the site visited and the time of the visit. For this type of data, clustering and classification can be used to both extract predefined and non-predefined clickstream behavior. Neural networks are useful if more complex patterns such as classifying the demographics of the users based on its clickstream or identifying the browsing behavior patterns need to be extracted from the data.

c. Methodologies

The previous section describes the more popular technologies in data mining that pertains to behavioral targeting. In this section, we will briefly discuss one of two of the better-known mining methodologies- Knowledge Discovery in Databases (KDD). CRISP-DM is also a notable process but since it has been commercialized by IBM into a tool, flexibility in mining abilities might be restricted. KDD is an academic notion but its steps can be followed with more improvisation. The actual implementation of data mining in the industry does not follow the exact steps defined in KDD but it is worthwhile to discuss them in light of how BT can be integrated. KDD is formally defined "as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Kurgan & Musilek, 2006). Although data mining is only a step in this process, the other steps can be critical perquisites to having good mined results.

KDD focuses on the overall process of discovering the data and storing the data, selecting the right target datasets, to the optimization of algorithms and learning methods that can be applied to large datasets, and how the results should be interpreted and visualized. In the article by Fayyad, Piatetsky-Shapiro, and Smyth, the authors highlight each step as follows:

- Data discovery- learning the data domain and the goal of the process or application.
- [Selection] Target data- select the data or focus on a subset of the data for the knowledge extraction.
- [Preprocessing] Preprocessing data- cleaning the data of extra noise or outliers and collecting information on data issues expected. If there are known causes of anomalies in the data, clean it first prior to continuing with KDD.
- [Transformation] Data reduction and projection- examine the variables and reduce the variable set to a minimum number of variables that can effectively reach a goal.
- [Data mining] Choose the function of data mining, whether it be classification or clustering or another technique. (purpose of the model)
- [Data Mining] Select the mining algorithm(s) that will determine how patterns will be searched, their appropriate parameters, and the preference criterion (a basis of preference of one model or set of parameters over another such as goodness-of-fit, over fitting the model).

- [Data Mining] Execute the searches for patterns defined by method(s) and algorithm(s) in previous steps.
- [Interpretation] Interpret the discovered patterns by business domain analysts; produce visualization of the results, or returning to previous steps to refine searches.
- [Knowledge] Incorporate the discovered knowledge into existing systems, taking initiative actions on new knowledge, or document and report results out.

Figure 1. Overview of the steps constituting the KDD process

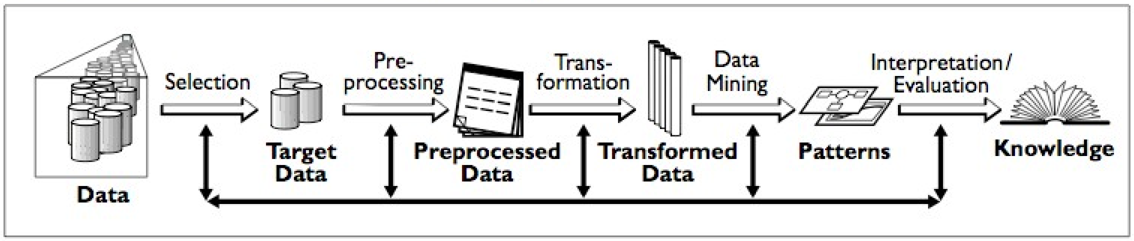


***Figure D8*- This diagram shows the basic steps of Knowledge Discovery in Database. Depending on the nature of the data, not all steps are required such as transformation.**

Reprint from "The KDD Process for Extracting useful Knowledge from Volumes of Data" by Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996.

The behavioral targeting adaptation of the KDD process does not deviate from the process model since KDD is designed for generic data mining. A detailed diagram of the KDD process model for BT, incorporating the data technologies mentioned in the survey is provided later in the survey.

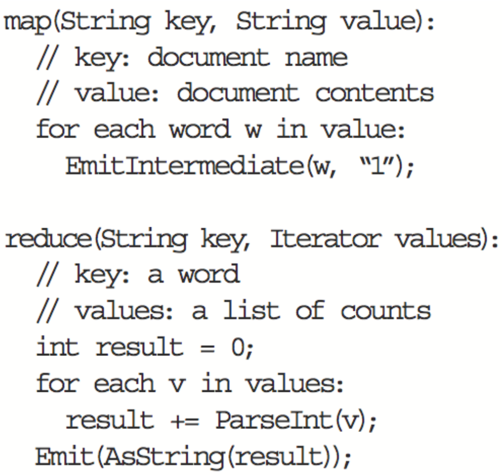
3. Processing Large Datasets

Aside from storage and knowledge extraction, processing of the large datasets and prepping the data before mining plays an important role in BT. Because of the need for high performance in terms of speed, processing big data requires parallelism working with cluster machines. Google conceived this idea and implemented a programming model, based on functional programming, which processes huge amounts of data according to a user-defined map function. Google pioneered the MapReduce framework and is now adopted by the open-source community, of which Hadoop is most well known and used by Amazon in its Elastic MapReduce cloud service.

To fully understand the notion of MapReduce, we will discuss the basics of Mapping and Reducing as a framework in this section. A high level view of how this can be implemented will be included in the summary of the data technologies mentioned in this paper.

MapReduce is a framework. Therefore, it is a set of libraries that controls parallel processing that exposes an API for implementing systems to call the Map and Reduce functions. It is not a developed user interface tool but Amazon does offer MapReduce as a toolset in the form of an online service.

MapReduce's key concepts are indicative of its name: Map and Reduce. Mappers are functions that take in a key-value pair as a parameter. The key-value pair can be of any primitive or complex data structure depending on what the function is trying to achieve. The purpose of a Map function is to try to map a key (a simple component/attribute of an input) to a value that fits the criteria of the map function found in the input. For each result found by a mapper, it is added to a collection. Similarly, a Reduce function does exactly what it's named for. The goal of Reducers is to combine all the values for a single key. A Reduce function will also take in a key-pair value. The output of a Reducer is arbitrary depending on what the goal of the reducer is. To illustrate this, we will extract an example presented in the article by Dean and Ghemawat, 2010. The example is the problem of finding the total number of occurrence for a word.



```
map(String key, String value):
  // key: document name
  // value: document contents
  for each word w in value:
    EmitIntermediate(w, "1");

reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
  int result = 0;
  for each v in values:
    result += ParseInt(v);
  Emit(AsString(result));
```

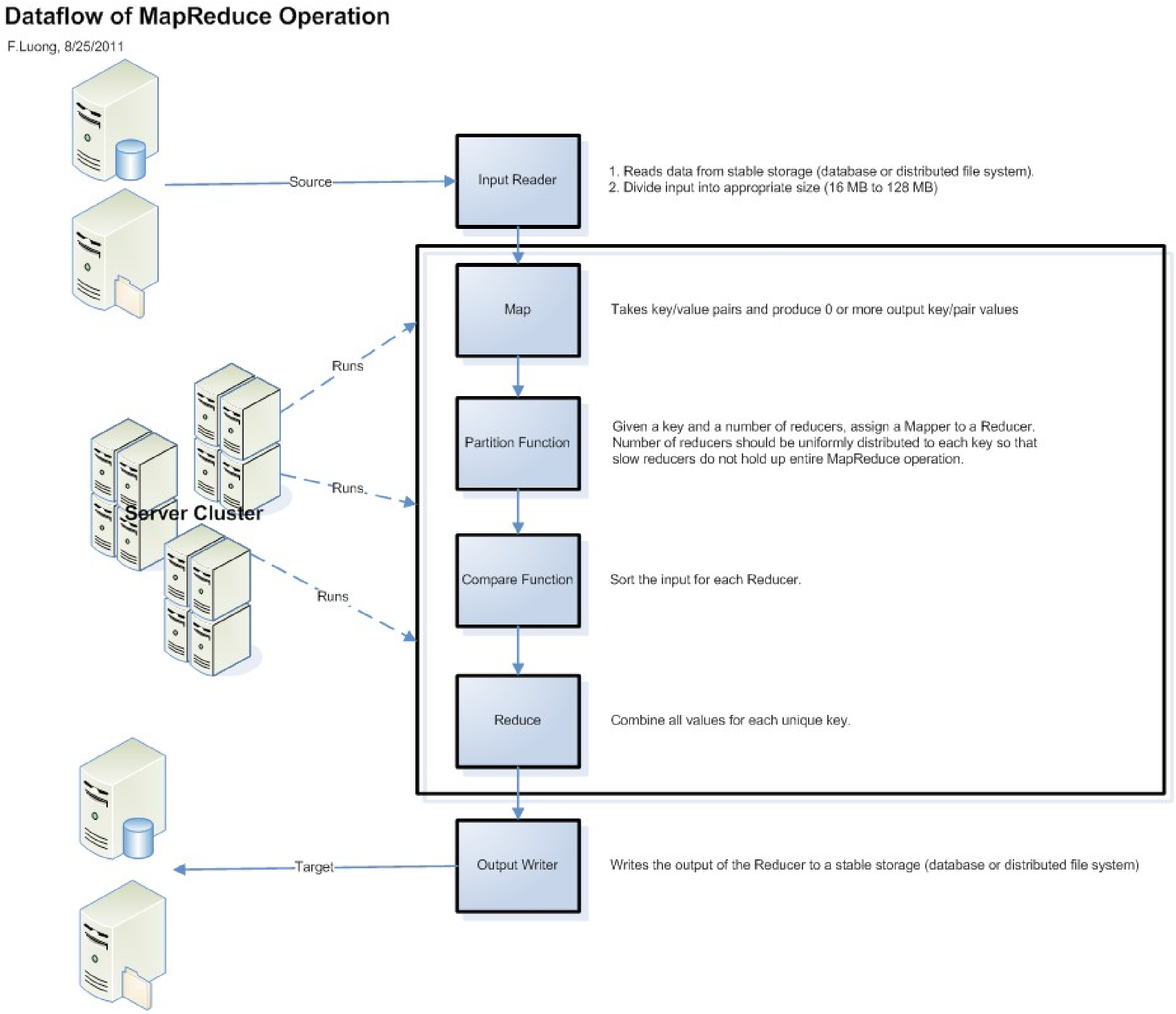
The map function takes a key-pair value of (document name, document content) and for each word in the document content, it tags a "1" for each occurrence found. The [word, "1"] pair is added to a collection each time it is found. Somewhere in the map function, even though not shown, will assign all "1" to a word in a data structure to prepare the result for the reduce function.

The reduce function will take a key-pair value of a (word, result from map for the word) and process it. In this case, the key is a word and the value (result from the map function) is a list of counts that pertains to the word (a list of "1"). The reduce function will add up all the "1" and come up with the final total of occurrence for the word.

***Figure D9*- Map and Reduce function for solving the problem of counting each word in a document. Reprinted from "MapReduce: A Flexible Data Processing Tool" by Jeffrey Dean and Sanjay Ghemawat.**

The logic of a map and reduce function for the example problem does not seem like a new innovation for solving "count" problems. However, the power of mapping and reducing lies not in the functions itself but with the addition of the parallel processing nature of this framework. Put into the context of having to process this problem for large amounts of documents with a large amount of words in each document, the performance of MapReduce is obvious.

The diagram below summarizes the flow of the framework.



Dataflow of MapReduce Operation

F.Luong, 8/25/2011

1. Reads data from stable storage (database or distributed file system).
2. Divide input into appropriate size (16 MB to 128 MB)

Takes key/value pairs and produce 0 or more output key/pair values

Given a key and a number of reducers, assign a Mapper to a Reducer. Number of reducers should be uniformly distributed to each key so that slow reducers do not hold up entire MapReduce operation.

Sort the input for each Reducer.

Combine all values for each unique key.

Writes the output of the Reducer to a stable storage (database or distributed file system)

***Figure D10*- This diagram shows the flow of the how the MapReduce operation is processed. The steps from Map to Reduce can be run be in parallel. The reading from source and writing to target, even though not shown in parallelism can be architected in parallel reads and writes with the right infrastructure. Interpreted from Wikipedia article.**

As indicated in the diagram, MapReduce can be used with heterogeneous systems for both the source and targets. With this flexibility, MapReduce framework provides a framework that can be adapted in most database platforms and file system. One of the benefits that MapReduce provide is the ability to process complex functions over a dataset. For behavioral targeting, the processing rules are limitless and unpredictable based on the goal of implementing BT. With MapReduce, custom functions can be created without the need to rely on SQL to process the rules. One constraint of MapReduce is that Reducers cannot start without all Mappers finishing which is why it is critical that workload is distributed evenly among Mappers and Reducers so that no part of the process becomes a bottleneck.

Summary of Big Data

The data technology for behavioral targeting covering its many different facets is an immense space of options, ranging from smaller scale of a single website to a larger advertising network trying to implement BT for multiple clients. The technologies and methodologies discussed in this paper do not represent all of the options available but are the commonly used and popular ones today. The table below summarizes the technologies and the area of interest it pertains with regard to BT.

Summary of Data Technology for Behavioral Targeting		
Area of Interest	Technology/Methodology	
Infrastructure	Parallel Processing in Machine clusters	
Data Storage	Google Bigtable	• Google File System • SSTable
	Amazon S3	
Data Mining	Methodology	• Knowledge Discovery in Databases (KDD)
	Model	• Clustering • Classification
	Algorithm	• Neural Network • Decision Tree • Association Rules
Data Processing	MapReduce	

***Table D2*- Summary of technologies and methodology for Behavioral Targeting.**

To summarize and put into perspective all the areas of interest and technologies/methodologies, a conceptual architecture view is presented below to show how behavioral targeting can be implemented.

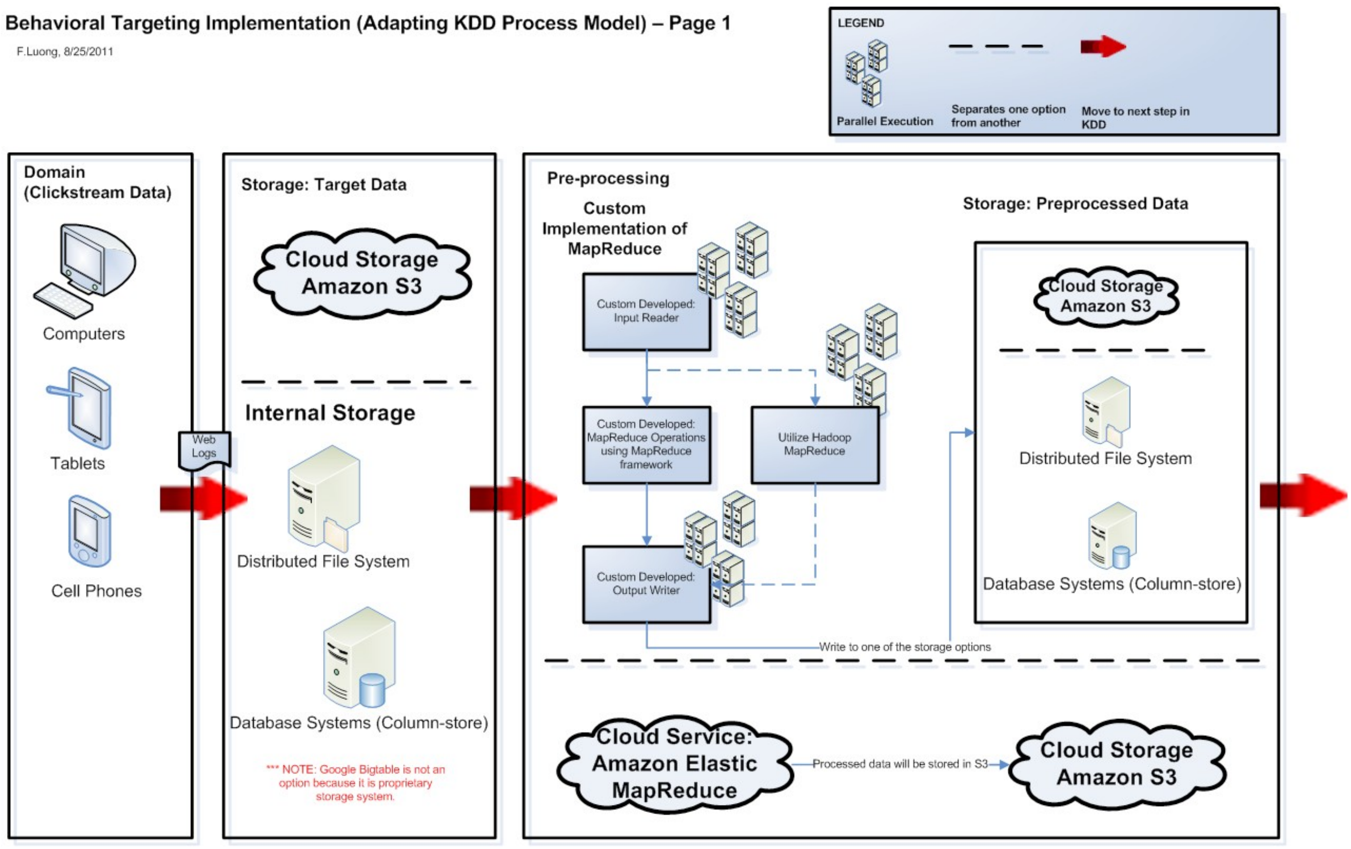
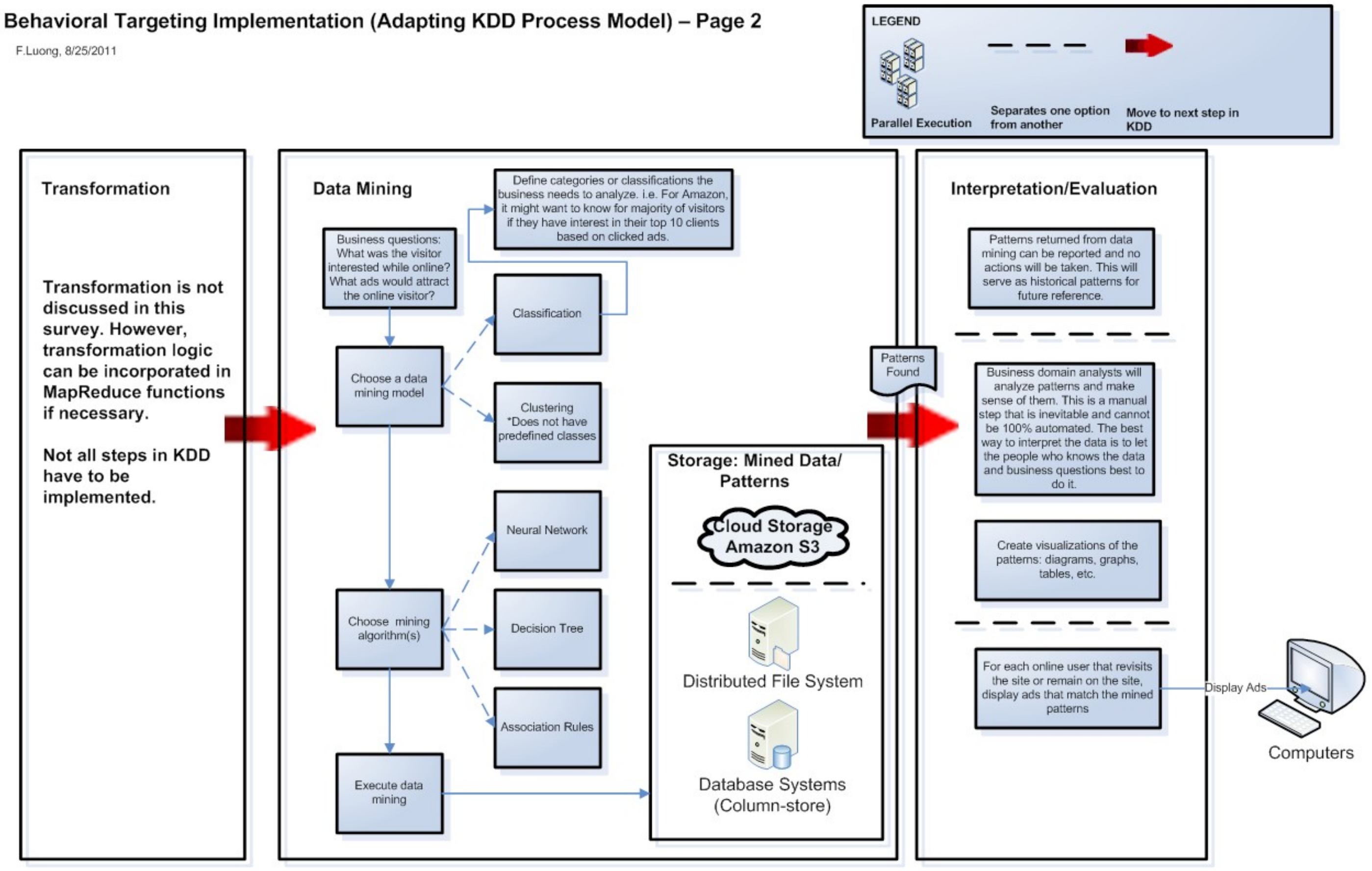

Behavioral Targeting Implementation (Adapting KDD Process Model) – Page 1
F.Luong, 8/25/2011
LEGEND
Parallel Execution
Separates one option from another
Move to next step in KDD
Domain (Clickstream Data)
Computers
Tablets
Cell Phones
Web Logs
Storage: Target Data
Cloud Storage Amazon S3
Internal Storage
Distributed File System
Database Systems (Column-store)
*** NOTE: Google Bigtable is not an option because it is proprietary storage system.
Pre-processing
Custom Implementation of MapReduce
Custom Developed: Input Reader
Custom Developed: MapReduce Operations using MapReduce framework
Utilize Hadoop MapReduce
Custom Developed: Output Writer
Write to one of the storage options
Storage: Preprocessed Data
Cloud Storage Amazon S3
Distributed File System
Database Systems (Column-store)
Cloud Service: Amazon Elastic MapReduce
Processed data will be stored in S3
Cloud Storage Amazon S3

***Figure D11*- High-level conceptual view of Behavioral Technology adaptation of KDD. This diagram also lists the technologies that are relevant in each step of the KDD process.**



Behavioral Targeting Implementation (Adapting KDD Process Model) – Page 2
F.Luong, 8/25/2011
LEGEND
Parallel Execution
Separates one option from another
Move to next step in KDD
Transformation
Data Mining
Business questions: What was the visitor interested while online? What ads would attract the online visitor?
Choose a data mining model
Classification
Clustering *Does not have predefined classes
Neural Network
Choose mining algorithm(s)
Decision Tree
Association Rules
Execute data mining
Storage: Mined Data/ Patterns
Cloud Storage Amazon S3
Distributed File System
Database Systems (Column-store)
Patterns Found
Interpretation/Evaluation
Create visualizations of the patterns: diagrams, graphs, tables, etc.
Display Ads
Computers

Transformation is not discussed in this survey. However, transformation logic can be incorporated in MapReduce functions if necessary.

Not all steps in KDD have to be implemented.

Define categories or classifications the business needs to analyze. i.e. For Amazon, it might want to know for majority of visitors if they have interest in their top 10 clients based on clicked ads.

Patterns returned from data mining can be reported and no actions will be taken. This will serve as historical patterns for future reference.

Business domain analysts will analyze patterns and make sense of them. This is a manual step that is inevitable and cannot be 100% automated. The best way to interpret the data is to let the people who knows the data and business questions best to do it.

For each online user that revisits the site or remain on the site, display ads that match the mined patterns

Figure D12- Continuation of Figure D11

Data Challenges and Future Advancements

Data is growing at an exponential rate. It is getting larger in numbers and in width (in terms of number of database columns). Even with today's powerful machines and state-of-the-art algorithms and mining software, dealing with data is still a challenging task and more advancement is expected in the future. One of the more prominent issues in data mining is the statistical inference that can be deducted from mining even huge amount of random data. Statisticians fear that when there is a large amount of data and you are continuously analyzing it, there is bound to be some patterns that can be extracted (Piatetsky, Matheus, Smyth, & Uthurusamy, 1994). However, these patterns might not have statistical relevance. In a way, this is creating false knowledge or knowledge that has no importance on the business question that is being asked at hand. There are ways to test the statistical relevance of the results by applying some controls such as the Bonferroni adjustments [1] or randomized testing procedures (Piatetsky et. al).

A similar issue, even with proper statistics, the plethora of data being analyzed can lead to an overabundance of useless and redundant patterns. A common approach to finding real patterns is to focus on obvious patterns that will not change throughout iterations of experiments. Once these real patterns are discovered, the task of distinguishing what patterns pertains to the business requires business domain knowledge. Thus, it is imperative that knowledgeable business analysts or data stewards are available during the knowledge discovery process.

The natures of the massive datasets also present a challenging problem. Businesses do not have perfect data due to many reasons. Data can be missing to answer business critical questions or it can be changing rapidly that invalidates previously discovered patterns. For non-stationary data, it is more difficult to keep up with the pattern searching so both business and IT will have to strategize on managing these data.

Aside from data itself, implementing KDD to work with existing applications is a technical challenge. The infrastructure and technology stack at various companies differ so the undertaking of KDD will most likely be of a significant effort. In some companies, important and enterprise data are housed in older technologies such as DB2 and mainframe databases. The mining techniques and algorithms will have to integrate with these systems.

Future advancements in KDD focus on multistrategy systems and large-scale databases. A single learning method does not solve all business questions or problems. Research has shown that no single method is best across a range of problems. Instead, multistrategy systems will “apply a number of different methods to the same task and select rules from the best method” (Piatetsky et. al). Large-scale databases enable better performance of processing large-scale data. They are unlike traditional relational database structure and mechanism. The more popular database systems for large-scale data includes parallel DBMS, key-value storage (which was discussed in map reduce in this paper), column-oriented, and streaming processing (focus on processing event-based stream of data) (“Database Technology for Large-Scale Data”, 2011).

[1] Bonferroni adjustment- The Bonferroni correction is a method used to counteract the problem of multiple comparisons, developed and introduced by Italian mathematician Carlo Emilio Bonferroni. The correction is based on the idea that if an experimenter is testing *n* dependent or independent hypotheses on a set of data, then one way of maintaining the family-wise error rate is to test each individual hypothesis at a statistical significance level of $1/n$ times what it would be if only one hypothesis were tested.

Issues

As discussed above, behavioral targeting (BT) is a sophisticated method of analyzing user data so that customized advertisements can be displayed to said user as they browse the Web. Many of us have already experienced the recommender style of behavioral targeting techniques as we shopped or browsed a particular website. Companies have realized the effects of exposing customers to items that are alike or complimentary to what they purchase. As stated by Howard Beales,

> "Advertising rates are significantly higher when BT is used…, advertising using BT is more successful than standard RON advertising, creating greater utility for consumers and clear appeal for advertisers because of the increased conversion of ads into sales" (p. 7).

However, there are advantages and disadvantages to this technology. Additionally, the legality of behavioral technology is yet to be explored. Some lawmakers have taken notice and this will become more prevalent as consumers are educated about the facts of data collection and the possible ramifications of its implementation. We will take a succinct look at the current effects of behavioral technology regulation in Europe and what we can expect from laws created in the United States. Lastly, we will examine the issues of this disruptive technology in an effort to gain knowledge of the impact on traditional marketing and advertising methods.

Advantages of this technology are quite obvious. It can be expedient to get exposure to an item that is related to or expands ones knowledge of a product. Similarly, such exposure can incite further exploration within a product family or a related category. An example of this is discussed in Waisberg's article in which a consumer who purchases a book in a certain genre is exposed to other authors or relevant titles that he may have not have otherwise known about. In a flawless world, this scenario can be applied to almost all of our purchasing endeavors. We would have product recommendation for just about any product. There are those among us who may view this as the perfect personal shopping tool; viewing customized marketing objectives could lessen the time to sort through potential products. Contrarily, there are those who are knowledgeable of the associated risks involved in such a technology and the disadvantages that may follow.

The most visible disadvantages surround the debate of privacy and the levels of exploitation that can be used by aggressive advertisers. In order to provide recommended products, companies must first acquire, analyze and use collected data to customize these recommendations. Data is bought and sold between third party companies that may not have solicited user consent to use said data. Additionally, data is also collected unbeknownst to a user by means of browser cookies, web crawlers and similar tools. Data is also gathered from activity and information that a consumer may post on a social networking site.

Simple items such as warranty submission for a purchased product or the common loyalty card issued by the grocery store collects information that is later used to target consumers. The possible privacy violations with these methods of collection are quite obvious. The mask of anonymity is slowly being removed from even the basic of activities.

Presently there are three classifications of consumers as listed by Goldfarb et al. They include consumers who do not want their information collected, those who are unaware that such data is being collected and those who are aware but may not be knowledgeable of the risks.
Authors Helft and Tanzina describe an occurrence of an online advertisement that followed consumer Julie Matlin. Ms. Matlin recounts her experience with the online shoe retailer Zappos and the advertisement that followed her throughout her browsing. Matlin describes clicking on a pair of shoes that she liked but decided not to purchase the shoes. She states, "for days or weeks, every site I went to seemed to be showing me ads for those shoes." She also adds "It is a pretty clever marketing tool. But it's a little creepy, especially if you don't know what's going on." Matlin's experience depicts two of the classifications mentioned by Goldfarb et al. We can deduce that Matlin was unhappy with the shoe advertisement following her from site to site. This is merely a simple example but we can clearly expand on this event by replacing the shoe with any other item that we browse for purchase. While the exploitation is subtle in the fact that it does not solicit user input, it is becoming a topic that will need some regulation in order to protect consumers. Additionally, Waisberg raised and astute supposition that not only does behavioral technology link users to otherwise unknown products, it can also have an unintended consequence of narrowing the consumers choices. Waisberg states that behavioral technology

> "reduces the information-content to which the buyer is exposed. He sees only things which his browsing habits justify… things that are likely to be similar to the ones he saw in the past."

In essence, the consumer is no longer exploring per se. He is only shown the objects that match his browsing and purchasing habits. Waisberg dubs this a "narrow world" in which the consumer would only see products that are relevant to those browsing habits. All other content is not shown and the user is left with an over customized view of products. Waisburg concludes that a "good recommender system should achieve not only the accuracy in its recommendations, but also novelty and serendipity." Waisburg cites Resnick and Varian in iterating this issue of narrowing a consumer's view in that it creates a division between the consumer and his peers. The core of this argument is that behavioral targeting and recommender systems limit the occurrence of serendipity and consumers may "miss the things that might" be of interest. Targeting browsing habits could narrow the advertised products viewed but quantifying the impact that this could have has yet to be discovered.

As for the legality of behavioral targeting, there are no laws governing this topic as yet. Technology moves faster than the laws that administer it. The

Federal Trade Commission (FTC) is the legal authority to inhibit “unfair and deceptive” (FTC) practices and this includes online transactions. However, as aptly put by Earp and Baumer, the FTC cannot regulate the entire Internet. The Gramm-Leach-Bliley Act, Health Insurance and Accountability Act (HIPAA) and the Children Online Privacy and Protection Act (COPPA) are enacted to protect consumers on the Web. However, their span is limited to specific scopes of banking, health records and protection of children. The initial attempt to protect consumers brought forth the Fair Credit Reporting Act which has been expanded to include online activities. The FTC’s introduction of the “notice-and-choice” and “harm-based” models (FTC p.7) has only been partially effective. While the former produced consent forms that were verbose and in some form incomprehensible, the latter lacked focus and was not expansive. While it is apparent that some businesses adhere to the notice, consent, access, security and redress policies proposed by the FTC, they are not sufficient at impacting behavioral technology. Thus, the FTC has proposed three new policies in an effort to engender change for data collection and targeting. The three policies are listed as “privacy by design” (FTC p.9), data choices for consumers and clear policies on data collection, sharing and storage.

The “privacy by design” (FTC) policy aims to educate businesses on the importance of focusing on the privacy of their consumers. Businesses should have clear plans on how data is collected, stored and shared. Additionally, businesses should invest in their privacy policy by educating their employees on the importance of adhering to the designed policies. The listed data and clear policies outlined by the FTC highlights the need for businesses to clearly post their privacy policies and give consumers the option to disagree with the posted privacy policy. While these are advancing steps in online consumer protection, they are early in effecting the targeted advertising practices that enables behavioral targeting in the United States.

The acts of behavioral targeting are viewed differently in the United Kingdom in that explicit directives have been enacted to protect consumers. Specifically, Directive 95/46/EC and Directive 2002/58/EC (Hustinx p.3) which are commonly known as the “Data Protection Directive” and the “e-Privacy Directive” respectively and have incited much debate. As stated by Hustinx, the two main elements are the “process of personal data” and the “storing or accessing of information stored in the user’s terminal” (Hustinx p.3). Hustinx highlights Article 5(3) of the e-Privacy Directive and defends its posture in protecting the once defined “opt-in” policy. Additionally, Hustinx cites Article 5(3)’s explicit phrasing which states

> “… user concerned has given his or her consent, having been provided with clear and comprehensive information…” (p.4)

In essence, the e-Privacy Directive makes it compulsory to first receive permission from the user in order to store or use information gathered. It goes

further to state that information provided to the user concerning collection of data must be “clear, precise and easily understandable.” (Hustinx p.5)
It is essential for the aforementioned directives to engender international support so that similar laws can be brought forth globally. This may seem impossible but one cannot ignore that we are all ubiquitously connected. The protection of consumers cannot be ignored in favor of profitability. Our privacy boundaries are being whittled away with each unchallenged intrusion into our daily activities. One can go as far as to say that we allow these impositions because the short term benefit appears favorable. However, an informed user is a powerful consumer. Thus, as consumers become more aware of the targeting activities that aggressive companies employ, one can hope that they will begin to question and investigate these acts. Hustinx says it best that the act of “tracking consumer behavior online is a highly intrusive practice” (Hustinx p.7) and should have more restrictions on the action.

What constitutes a reasonable expectation of privacy? There is no one answer to this question. As access to information becomes more pervasive, it is inevitable that aspects of privacy will decrease. However, it would be an ignominious failure to society if some form of regulation is not introduced to protect the collection and sharing of private data. It would be unwise to completely inhibit data collection and processing; but there must be measures in place to regulate the process. The fact remains that technology evolves faster than the laws that govern it. The advancements in behavioral technology is rather Janus faced in that it introduces consumers to products that fit their habits but intrudes and ignores the privacy boundaries that are inherent. Must one accept the former and give in to the latter? How much are we willing to share in order to get this customized approach? Behavioral technology has passed the infancy stage and is now at the refinement stage; it permeates the Internet. The figures below depict privacy maps for the United States and the European Union.

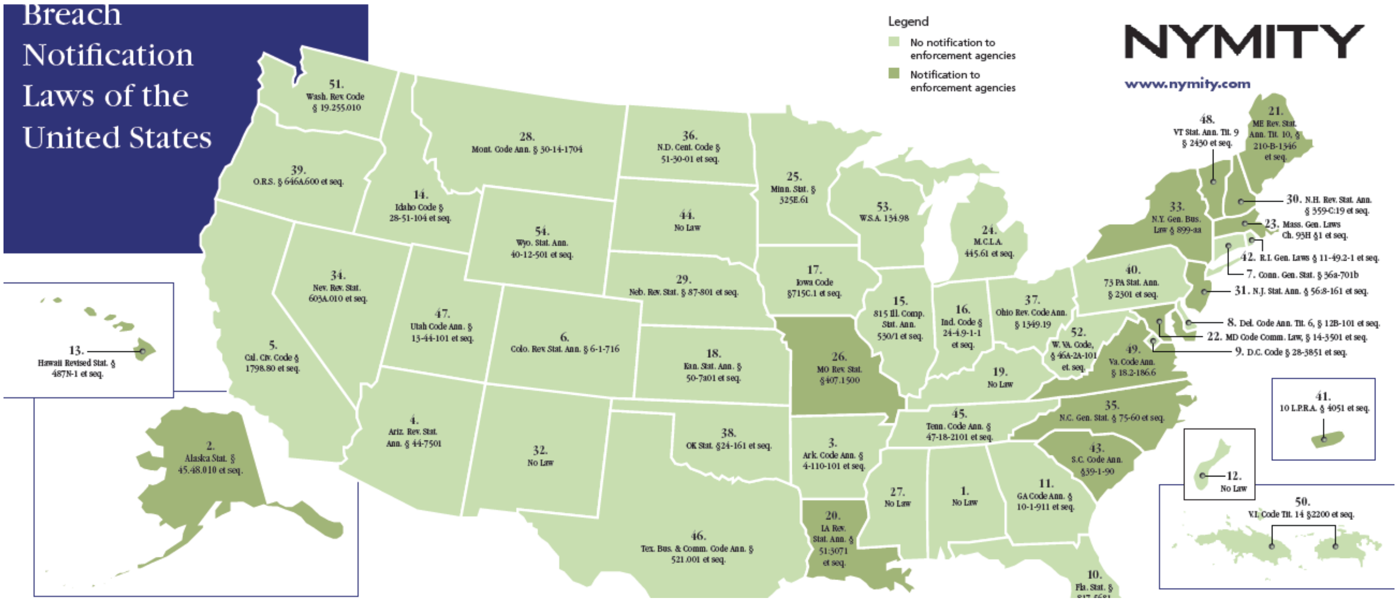

Breach
Notification
Laws of the
United States
Legend
No notification to enforcement agencies
Notification to enforcement agencies
NYMITY
www.nymity.com
51. Wash. Rev. Code § 19.255.010
39. O.R.S. § 646A.600 et seq.
14. Idaho Code § 28-51-104 et seq.
28. Mont. Code Ann. § 30-14-1704
54. Wyo. Stat. Ann. 40-12-501 et seq.
36. N.D. Cent. Code § 51-30-01 et seq.
44. No Law
29. Neb. Rev. Stat. § 87-801 et seq.
25. Minn. Stat. § 325E.61
53. W.S.A. 134.98
24. M.C.L.A. 445.61 et seq.
17. Iowa Code §715C.1 et seq.
15. 815 Ill. Comp. Stat. Ann. 530/1 et seq.
16. Ind. Code § 24-4.9-1-1 et seq.
37. Ohio Rev. Code Ann. § 1349.19
40. 73 PA Stat. Ann. § 2301 et seq.
33. N.Y. Gen. Bus. Law § 899-aa
48. VT Stat. Ann. Tit. 9 § 2430 et seq.
21. ME Rev. Stat. Ann. Tit. 10, § 210-B-1346 et seq.
30. N.H. Rev. Stat. Ann. § 359-C:19 et seq.
23. Mass. Gen. Laws Ch. 93H §1 et seq.
42. R.I. Gen. Laws § 11-49.2-1 et seq.
7. Conn. Gen. Stat. § 36a-701b
31. N.J. Stat. Ann. § 56:8-161 et seq.
8. Del. Code Ann. Tit. 6, § 12B-101 et seq.
22. MD Code Comm. Law, § 14-3501 et seq.
9. D.C. Code § 28-3851 et seq.
52. W. VA. Code, § 46A-2A-101 et. seq.
49. Va. Code Ann. § 18.2-186.6
13. Hawaii Revised Stat. § 487N-1 et seq.
2. Alaska Stat. § 45.48.010 et seq.
5. Cal. Civ. Code § 1798.80 et seq.
34. Nev. Rev. Stat. 603A.010 et seq.
47. Utah Code Ann. § 13-44-101 et seq.
6. Colo. Rev Stat. Ann. § 6-1-716
18. Kan. Stat. Ann. § 50-7a01 et seq.
26. MO Rev. Stat. §407.1500
19. No Law
4. Ariz. Rev. Stat. Ann. § 44-7501
32. No Law
38. OK Stat. §24-161 et seq.
3. Ark. Code Ann. § 4-110-101 et seq.
45. Tenn. Code Ann. § 47-18-2101 et seq.
35. N.C. Gen. Stat. § 75-60 et seq.
41. 10 L.P.R.A. § 4051 et seq.
43. S.C. Code Ann. §39-1-90
12. No Law
50. V.I. Code Tit. 14 §2200 et seq.
46. Tex. Bus. & Comm. Code Ann. § 521.001 et seq.
20. LA Rev. Stat. Ann. § 51:3071 et seq.
27. No Law
1. No Law
11. GA Code Ann. § 10-1-911 et seq.
10. Fla. Stat. §

***Figure F1*: Breach Notification Laws of the United States adapted from Privacy by Design, Ann Cavoukian. http://www.privacybydesign.ca/content/uploads/2010/03/NYMITY-USA_map.pdf**

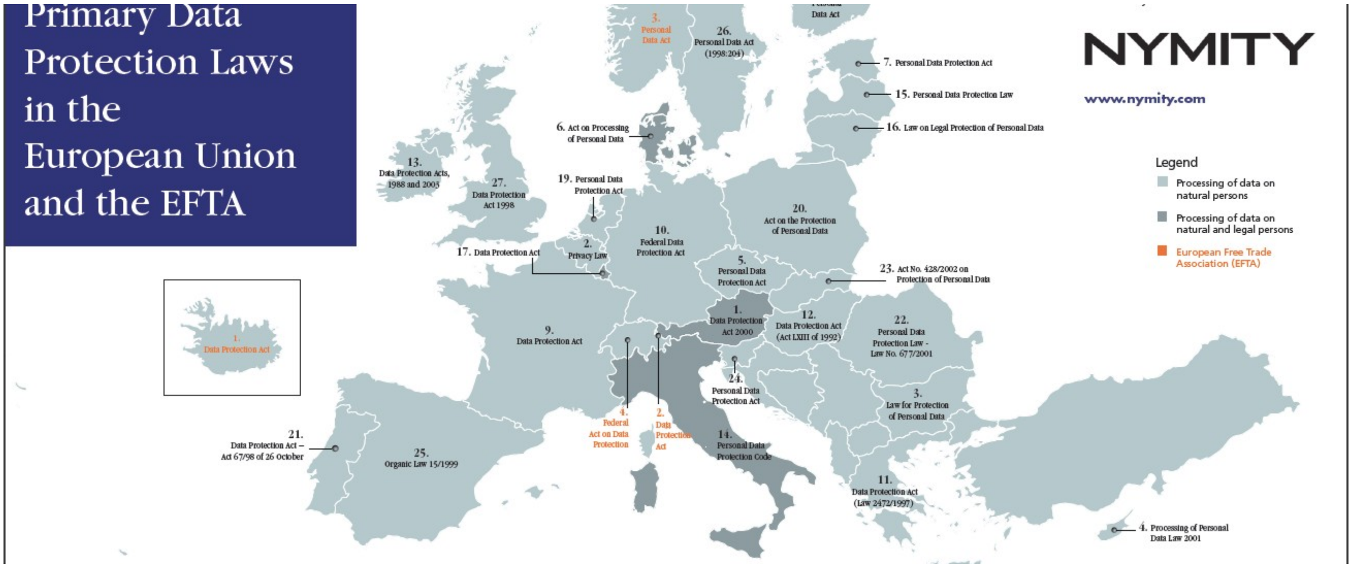

Primary Data Protection Laws in the European Union and the EFTA
NYMITY
www.nymity.com
Legend
Processing of data on natural persons
Processing of data on natural and legal persons
European Free Trade Association (EFTA)
Data Act
3. Personal Data Act
26. Personal Data Act (1998:204)
7. Personal Data Protection Act
15. Personal Data Protection Law
16. Law on Legal Protection of Personal Data
6. Act on Processing of Personal Data
13. Data Protection Acts, 1988 and 2003
27. Data Protection Act 1998
19. Personal Data Protection Act
20. Act on the Protection of Personal Data
10. Federal Data Protection Act
2. Privacy Law
17. Data Protection Act
5. Personal Data Protection Act
23. Act No. 428/2002 on Protection of Personal Data
1. Data Protection Act
1. Data Protection Act 2000
12. Data Protection Act (Act LXIII of 1992)
22. Personal Data Protection Law - Law No. 677/2001
9. Data Protection Act
24. Personal Data Protection Act
3. Law for Protection of Personal Data
4. Federal Act on Data Protection
2. Data Protection Act
14. Personal Data Protection Code
21. Data Protection Act – Act 67/98 of 26 October
25. Organic Law 15/1999
11. Data Protection Act (Law 2472/1997)
4. Processing of Personal Data Law 2001

***Figure F2*: Data Protection Laws in the European Union adapted from Privacy by Design, Ann Cavoukian. http://www.privacybydesign.ca/content/uploads/2010/03/NYMITY-EU_map.pdf**

It is immediately noticed that privacy coverage spawns wider in the European Union than in the United States. While the European Union stringently enforces their “opt-in” policies, the United States takes on an “opt-out” policy. There are great differences in the way data collection activities precede between the two policies. While the former places the burden on the business to comply with data privacy directives, the latter places the burden on the consumer to “opt-out” if so desired. Both policies attempt to provide notice to consumers of data collection and sharing activities. However, the European Union has taken a stronger stance on the matter. The United States has to catch-up to the policies currently enacted in the European Union.

References

Alves, R., Belo, O. “Mining clickstream-based data cubes”. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.358&rep=rep1&type=pdf.

Beales,H. (2010). The Value of Behavioral Targeting. Retrieved from http://www.socialized.fr/wp-content/uploads/2010/08/beales-etude-sur-le-ciblage-comportemental-sur-internet-ppc4bible.pdf

Canny, J., Zhong, S., Gaffney, S., Brower, C., Berkhin, P, John, G.H. (2011). US Patent No. 7,921,069 Washington, D.C. US Patent & Trademark Office.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M.,… Gruber, R.E. Bigtable: A distributed storage system for structured data”. (2006). OSDI'06: Seventh Symposium on Operating System Design and Implementation. Seattle, WA. Retrieved from http://labs.google.com/papers/bigtable-osdi06.pdf.

Chen, J., Nairn, R., & Chi, E. 2011. Speak little and well: recommending conversations in online social streams. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI '11). ACM, New York, NY, USA, 217-226. DOI=10.1145/1978942.1978974 http://doi.acm.org/10.1145/1978942.1978974

Chen, Y., Pavlov, D., Canny, J. (2009). Large-scale behavioral targeting. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Presented at the Knowledge Discovery and Data Mining, ACM, New York, NY, USA. doi:10.1145/1557019.1557048

Chen, Y., Pavlov, D., Berkhin, P., Seetharaman, A., Meltzer, A. Practical Lessons of Data Mining at Yahoo!. (2009). CIKM’09. Hong Kong, China. Retrieved from ACM 978-1-60558-512-3/09/11.

Chen, Y.,Pavlov, D.,Berkhin, P., Canny, J.. (2010). Large-Scale Behavioral Targeting for Advertising over a Network- Google Patents. California. Retrieved from http://bit.ly/qC32Mk

Database technology for Large-Scale data”. Cubrid Official Blog. Retrieved from http://blog.cubrid.org/web-2-0/database-technology-for-large-scale-data/.

Dean, J., Ghemawat, S. MapReduce: A flexible data processing tool". (2010). *Communications of the ACM,* 53(1). Retrieved from http://cacm.acm.org/magazines/2010/1/55744-mapreduce-a-flexible-data-processing-tool/fulltext.

Doubleclick. Wikipedia. Retrieved from http://en.wikipedia.org/wiki/DoubleClick.

Earp, E.B., & Baumer, D. (2003). Innovative Web Use To Learn About Consumer Behavior and Online Privacy. *Communications of the ACM*, 46. Retrieved from http://web.sau.edu/lilliskevinm/csci660/2008Fall/papers/EarpBaumer.pdf

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. From data mining to knowledge discovery in databases". (1996). *AI Magazine*, 17(3). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.42.1071&rep=rep1&type=pdf.

Federal Trade Commission (FTC). (2010). *Protecting Consumer Privacy in an Era of Rapid Change*. Retrieved from http://www.ftc.gov/os/2010/12/101201privacyreport.pdf

Gannon, Dennis. "A computational data center- A science cloud". Retrieved from http://www.google.com/url?sa=t&source=web&cd=2&ved=0CCAQFjAB&url=http%3A%2F%2Fcyberaide.googlecode.com%2Fsvn%2Ftrunk%2Fmisc%2Fcloud-papers%2Fscience-data-center.pdf&rct=j&q=A%20Computational%20Data%20Center%20A%20Science%20Cloud&ei=4yw6To64O8eWtweLkLzrAg&usg=AFQjCNFs6EQ8583VOREsG9Ey7B19EPvHmQ&sig2=Hfo0xrrVmVidSLwVJ_qK-g&cad=rja.

Gerster, D., Awadaliah, A, Thampy, S. (2010). FORECASTING ASSOCIATION RULES ACROSS USER ENGAGEMENT LEVELS - Google Patents. Sunnyvale, CA. Retrieved from http://bit.ly/rfc6V5

Goldfrab, A., Tucker, C. (2011). Economic and Business Dimensions: Online Advertising; Behavioral Targeting and Privacy. *Viewpoints Communications of the ACM 54(5).*

Google begins Behavioral Targeting ad program" *Electronic Frontier Foundation*. Retrieved from http://www.eff.org/deeplinks/2009/03/google-begins-behavioral-targeting-ad-program.

Guo, Q., Agichtein, E., Clarke, C., and Ashkan, A. 2009. In the Mood to Click? Towards Inferring Receptiveness to Search Advertising. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01* (WI-IAT '09), Vol. 1. IEEE Computer Society, Washington, DC, USA, 319-324. DOI=10.1109/WI-IAT.2009.368 http://dx.doi.org/10.1109/WI-IAT.2009.368

Hustinx, P. “Do Not Track or Right to Track? – The Privacy implications of online behavioral advertising.” University of Edinburgh, School of Law. Edinburgh, 07 Jul. 2011. Retrieved from http://www.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Publications/Speeches/2011/11-07-07_Speech_Edinburgh_EN.pdf

Java, A., Song, X., Finin, T., et al. 2007. Why we twitter: Understanding microblogging usage and communities. Proc. WebKDD '07, 56-65.

The KDD process for extracting useful knowledge from volumes of data. (1996). *Communications of the ACM*, 39(11). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.2315&rep=rep1&type=pdf.

Kurgan, L.A., Musilek, P. A survey of knowledge discovery and data mining process models”. *The Knowledge Engineering Review*, 2(11), 1-24. Cambridge University Press. 2006. Retrieved from http://biomine.ece.ualberta.ca/papers/KER-KDDM2006.pdf.

Ngai, E.W.T., Xiu, L., Chau, D.C.K. Application of data mining techniques in customer relationship management: A literature review and classification”. (2009). *Expert Systems with Applications*, 36, 2592-2602. Retrieved from http://cjou.im.tku.edu.tw/bi2009/DM-usage.pdf.

Palankar, M., Onibokun, A., Iamnitchi, A., Ripeanu, M. Amazon S3 for science grids: a viable solution?” Retrieved from www.cse.usf.edu/~anda/papers/AmazonS3_TR.pdf.

Piatetsky-Shapiro, G., Matheus, C., Smyth, P., Uthurusamy, R. KDD-93: progress and challenges in knowledge discovery in databases”. (Spring 1994). *The American Association for Artificial Intelligence*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.30.375&rep=rep1&type=pdf.

Privacy by Design.(2011). Retrieved August 17, 2011 http://www.privacybydesign.ca

Rodden, K., Fu, X., Aula, A. and Spiro, I. Eye-mouse coordination patterns on web search results pages. In Proc. of CHI, 2008.

Rygielski, C., Wang, J.C., Yen, D.C. “Data mining techniques for customer relationship management”. (2002). *Technology in Society*, 24, 483-502. Retrieved from http://chern.ie.nthu.edu.tw/IEEM7103/923834-paper-1-june21.pdf.

Waisberg, I. Unintended Consequences of Targeting: Less Information, Less Serendipity Retrieved on 17 July, 2011 from http://online-behavior.com/targeting/unintended-consequences-of-targeting-part-ii-1484

Yahoo!’s Behavioral Targeting Advertising Solution. (2011). *Behavioral Targeting*. Retrieved August 27, 2011, from http://advertising.yahoo.com/products-solutions/behavioral-targeting.html