Data Mining and Audience Intelligence for Advertising
Manika Gupta
MCA V Sem, Jagan Institute of Management Studies, Delhi
Email: manikagupta31@gmail.com

ABSTRACT
Growth in the global advertising industry - especially the recent rapid growth in online advertising - has generated large volumes of data, bringing along with it many challenging data mining problems. Researchers from various disciplines have brought their expertise to solve these exciting problems, leading to a plethora of novel applications and new algorithms. Data mining techniques, which are to mine patterns and knowledge from large scale data, can provide effective solutions to the above problems in different types of online advertising, including sponsored search, contextual advertising, behavior targeting and so on.

INTRODUCTION
Global advertising is projected to exceed half-a-trillion dollars by the year 2010 [1]. Although online advertising is currently only a small part of this large enterprise, it is growing at a rapid pace [2]. The past few years have seen a tremendous growth in online advertising. Especially, the last two years have seen significant changes in the advertising industry both in terms of business deals as well as new industry initiatives. The explosion in the number of participants in the online advertising marketplace has generated large volumes of data and exciting data mining problems. Earlier research on search logs, web pages, social network and blogs had focused on information organization, retrieval and understanding [3]. Recently there has been strong research interest in the advertisement angle to all these information sources. Researchers have tackled several challenging problems on online monetization like sponsored search [4], contextual advertising for web pages [5], understanding user intent [6] and user demographics [7] for advertisements, mining user reviews for product pricing [8], predicting click-through rates for ads [9], just to name a few. Further, the on-line and offline advertising worlds are fast converging. For example digital marketplaces are migrating from the online world to TV and radio [10], and audience understanding work from offline media is trickling into the online realm.

Data mining researchers and practitioners in all these areas come from different communities, so there is strong need for single forum to bring together people involved in all aspects of digital advertising. The goal of this workshop was not only to increase communication between researchers working on seemingly different pieces of the advertisement pie, but to encourage data mining researchers to bring new ideas from related areas to solve the numerous challenges faced by the rapidly changing digital advertising industry. It brought together theorists, social network researchers, natural language researchers, information retrieval experts, audience understanding researchers, television advertisement analysts and many others, to promote a fruitful exchange of ideas to advance the field.

Online advertising is a complicated ecosystem, which involves multiple players, including advertisers, publishers, end users and many others. Delivering the right marketing messages to the right users in the right time is of great importance to the success of this ecosystem. To achieve this goal, the advertisements from advertisers, content from publishers and intents of the end users are to be understood intelligently.

1.1 Relevance of data mining to a business process
For data mining to impact a business, it needs to have relevance to the underlying business process. Data mining is part of a much larger series of steps that takes place between a company and its customers.

Take product marketing as an example. A marketing manager's job is to understand their market. With this understanding comes the ability to interact with customers in this market, using a number of channels. This involves a number of areas, including direct marketing, print advertising, telemarketing, and radio/television advertising, among others.

The issue that must be addressed is that the results of data mining are different from other data-driven business processes. In most standard interactions with customer data, nearly all of the results presented to the user are things that they knew existed in the database already. A report showing the breakdown of sales by product line and region is straightforward for the user to understand because they intuitively know that this kind of information already exists in the database.

Data mining, on the other hand, extracts information from a database that the user did not know existed. Relationships between variables and customer behaviors that are non-intuitive are the jewels that data mining hopes to find.
This is where interaction and context comes in. Marketing users need to understand the results of data mining before they can put them into actions. Because data mining usually involves extracting "hidden" patterns of customer behavior, the understanding process can get a bit complicated.

How does someone actually use the output of data mining? The simplest way is to leave the output in the form of a black box. If they take the black box and score a database, they can get a list of customers to target (send them a catalog, increase their credit limit, etc.). There's not much for the user to do other than sit back and watch the envelopes go out. This can be a very effective approach. Mailing costs can often be reduced by an order of magnitude without significantly reducing the response rate.

The output of data mining must be understood qualitatively. The user needs to view the output of the data mining in a context they understand. If they can understand what has been discovered, they will trust it and put it into use.

There are two parts to this problem:
1) Presenting the output of the data mining process in a meaningful way, and
2) Allowing the user to interact with the output so that simple questions can be answered.

Response rates and (probably most importantly) financial indicators (for example, profit, cost, and return on investment) give the user a sense of context that can quickly ground the results in reality.

1.2 No matter the size of business operation, data is everywhere.

When someone types keywords into a search engine and arrives at organization’s website, a record of that event can be captured and stored to inform the keywords users will target in the pay per click advertising campaigns. When a customer calls for support or to look into buying a product, the customer service agent usually logs the information electronically or on paper for future reference.

Data is everywhere, but the question is this: are you using it as effectively as possible? Most data is unstructured, meaning it isn’t consistently created or stored, or it can’t be easily used to inform your business decisions. It can be helped by analyzing the existing data and interpreting it or optimizing it to meet the business needs.

2. SOCIAL CHALLENGES IN ADVERTISING

Much of advertisement today is focused on audience intelligence - on what the advertisers and the data mining researchers know about the user. It is very important for the practitioners to be aware of the converse of audience intelligence - the intelligence of the audience.

More specifically we want to know: what do the users know when it comes to data mining and data use?. It is centered around this dilemma of social acceptance of targeted advertisements.

2.1 ISSUES:

- The challenges that public knowledge (or the lack of it) and perceptions raise for database advertisers and the intermediaries (such as Microsoft, Yahoo and Google).
- Societies perception using several national surveys of the internet using public opinions.
- The changes are tracked in targeted marketing from the time it was practiced by the street corner grocer, to the current practices by retailing and media regimes who use logic based on data mining.
- As the data gets bigger and technology begins to play a bigger role, the trend in business practice is moving towards identifying and targeting niche groups.
- The surveys showed that although people suspect that their data is being used (e.g. about 80% of users know that companies can track their behavior across web sites), they don't know the rules of the marketplace when it comes to using this data (e.g. 64% did not realize that supermarkets are allowed to sell other companies information on what they buy) [11].
- They are also bothered by the idea of “price discrimination” i.e. they do not like that that someone else can be targeted to pay less than them for the same product.
- The lack of transparency makes them feel vulnerable and encourages suspicion and anger at marketers, media and even the government (only 35% trust the government to protect them from marketers who misuse their information [12]).

2.2 SOLUTIONS:

The industry should promote transparency (increased user control over data), limit the amount/time information stored, and solicit data i.e. find out what a person wants to be and how would they like to be treated.

Data miners are referred as the new “story tellers" of the society - people who will soon have enough data to track the entire life of individuals.

Data miners should take this responsibility seriously and help users reach their aspirations.

3. MAJOR ISSUES IN ADVERTISING.

The first covers advertising on web pages, also known as contextual advertising.

The second focuses on sponsored search.

The third is on technologies that drive advertising on Web 2.0.

The last topic focuses on using techniques and principles from online advertising to content from other media such as TV.
3.1 Understanding Content & User Attention on a Web Page

This section focuses on technologies for advertising on web pages - content filtering and targeting. Contextual advertising places advertisements within the content of a generic web page. In order to maximize revenue while improving user experience, it is crucially important to place relevant advertisements to the page content.

3.1.1 One of the problems in advertising is to decide what to show and when to show it, so that the user's attention is best captured. The first topic in this section focuses on this interplay between user attention and information on a web page. The basic hypothesis is that in an information-rich environment, there is a constant competition for a user's attention.

Solution:
Treating the user's attention as a limited resource, the problem can be modeled as a dual-speed restless problem. An automatic mechanism that generates the most relevant information should be presented to the users. The proposed solution guarantees to maximize the users' total expected utility from the information they receive.

3.1.2 The core problem of content-targeted advertisement is to accurately match the content of a page to the sparse content of advertisements.

Solution:
The second topic in this section addresses this issue. The system assumes that each web page has been provided with several ad candidates of varying quality. Their goal is to re-rank the candidates so that the best ads are at the top of the list. A language independent machine learning system is present that re-ranks ad candidates based on a noisy-channel model using machine translation technologies. The system is validated through the experiments on a large number of advertisements appearing on real web pages.

3.1.3 The last topic in this section addresses an important problem in online advertising - filtering out unwanted content. Specifically, it is to detect whether a publisher web page contains content that is inappropriate for showing advertisement(s) on it. It can be referred as "Sensitive Webpage Classification for Content Advertising." A classification-based approach can be provided to this problem.

Solution:
First, a hierarchical sensitive content taxonomy should be used to customize this task. Next, outline an iterative training procedure based on active learning to learn from a large collection of labeled and unlabeled data. Finally, present experimental results to compare the performance of classifiers like SVMs and logistic regression on this task.

3.2 Sponsored search

The second section addresses two diverse problems in the practice of placing advertisements next to the search results (also known as "sponsored search"). It is to place advertisements on result pages from a web search engine. Advertiser-provided content, bidding keywords, relevance verification are the basic elements of sponsored search.

3.2.1 The first topic explores an alternative to the "pay-per-click" model that is popular today. Many challenges are involved in designing a "pay-per-action" business model, where payment is made based on user conversion or purchase. Although such a model is a natural progression from the two dominant business models: the pay per-impression model and the pay-per-click model. It has been discussed often in the advertising industry, but is not widely used yet.

The challenges faced in implementing such a system:
The challenge of encouraging advertisers to invest in tracking and disclosing conversions truthfully to the auctioneer mechanisms.

Solution:
When the conversion rate becomes part of the mechanism to rank ads, advertisers with higher conversion rates tend to be ranked higher than others with the same bid, and this encourages truthful reporting.

3.2.2 The next area deals with maximizing advertiser return on investment (ROI), specifically about how to measure the effectiveness of methods implemented to increase ROI. For example, online search systems that display ads features that advertisers can use to fine-tune and enhance their ad campaigns. Thus, for both sponsored search and contextual advertising, it is critical to understand the users' need, which can be inferred from the users' queries and demographics.

3.3 Sentiment Classification, Opinion Mining and Analyzing Information Flow in Blogs

Social aspects of advertising are becoming increasingly important. Web 2.0 sites like blogs and review sites are important areas for targeted advertisement. Specifically, it is to detect whether a publisher web page contains content that is inappropriate for showing advertisement(s) on it. It can be referred as "Sensitive Webpage Classification for Content Advertising." A classification-based approach can be provided to this problem.

Solution:
First, a hierarchical sensitive content taxonomy should be used to customize this task. Next, outline an iterative training procedure based on active learning to learn from a large collection of labeled and unlabeled data. Finally, present experimental results to compare the performance of classifiers like SVMs and logistic regression on this task.
kernels are similarity metrics in non-Euclidean information spaces, which have been found to produce state of the art results for document classification.

3.3.1 This addresses problem of **sentiment mining** - to find the object of the sentiment or opinion \[^{14}\]. For example, a review blog on a digital camera may praise one feature, while panning another. How is an advertiser to know automatically, which features are liked by the reviewers and which are not? A rule-based approach is proposed to extracting topics from opinion sentences by assuming the sentences are identified from texts in advance. They build a sentiment dictionary and define several rules based on the syntactic roles of words.

3.3.2 The last topic in this section is “Discovering Information Diffusion Paths from Blogosphere for Online Advertising” by Avare Stewart, Ling Chen, and Raluca Păiu Wolfgang Nejdl \[^{15}\] from L3S, Germany. The authors propose to discover information diffusion paths from the blogosphere to track how information travels from blog to blog. This is useful for determining the most effective places to advertise in the blog world. Their method first analyzes the content of blogs to detect trackable topics. Then they model a blog community as a blog sequence database, and formalize and solve the problem of discovering diffusion paths as one of frequent pattern mining. Experiments conducted on real life dataset show that their algorithm can discover the information diffusion paths efficiently.

3.4 TV and Other Broadcast Content, and their Relation to Online Advertisement

The final section focused on two diverse topics.

3.4.1 The first is about **targeting online transcripts of TV news**. Traditional content advertising has been targeted for webpages, and has relied on keyword extraction \[^{16}\]. However, the existing technologies cannot be easily applied to find keywords from online broadcasting content, which usually contain more specific phrases and wordings in certain communities than in general Web-page content. “Finding Keyword from Online Broadcasting Content for Targeted Advertising” presents a sequential pattern mining-based method to discover language patterns from online broadcasting content. Starting with selected keyword seeds, and by iteratively applying the language pattern mining and keyword extraction steps, the proposed technique avoids any tedious labeling work for this task. Experiments on some real-world data show that the proposed keyword extraction algorithm can significantly outperforms some baseline methods.

3.4.2 The second and last topic in this section is “From TV to Online Advertising: Recent Experience from the Spanish Media". The advance of the Internet as a competitor with traditional media (radio, TV, newspapers and magazines) has attracted many advertisements. However, the traditional analytical tools for media planning may not be directly applicable on online advertising.

4. ASPECTS IN ONLINE ADVERTISING

4.1 Online Effects of Offline Ads
Online advertising and offline ads seem to be well separated. However, their impact on users' daily life is hard to distinguish. Clearly, online advertising can affect users' online behaviors and vice versa. Diane Lambert and Daryl Pregibon's paper “Online Effects of Offline Ads” proposes a methodology for assessing how ad campaigns in offline media such as print, audio and TV affect online interest in the advertisers brand. Online interest can be measured by daily counts of the number of search queries that contain brand related keywords, by the number of visitors to the advertisers web pages, by the number of pageviews at the advertisers websites, or by the total duration of visits to the advertisers website.

4.2 Sponsored Ad Based Similarity
This presents a method for mining the intelligence of advertisers to detect product similarities and generate accurate recommendations. The basic assumption is that if object A and object B each lead to the display of sponsored ad C, then this is an indication of similarity between A and B. A general framework for leveraging linked advertisements to detect object similarity is proposed.

4.3 Personalized Online Commercial Intention
Understanding users' intention, especially their online commercial intention through their search queries is very important to online advertising. It can help search engines provide proper search results and advertisements; help Web users obtain the right information they desire; and help the advertisers make revenue from the potential transactions.

5 DATA MINING ALGORITHMS

5.1 Uncertain Data Mining
In many real applications, the records are not specified precisely, and may have errors built into them. In such cases, one may have to work with the errors in order to obtain more accurate data mining results. For example, if an attribute has a larger error for a classification problem, then probably another attribute with lower uncertainty can be used, even though it may not have a very high predictive power. A density based mining of uncertain data mining problems has been proposed.

5.2 High Dimensional Data Mining:
The problem of high dimensional data mining has been widely studied by the research community and has continued to be an
intriguing problem over many years. Two important issues often arise in many high dimensional data mining applications: (1) Performance issues because of data sparsity (2) Qualitative issues because of the lack of contrast in distance calculations. For example, if a small number of points in high dimensional space are uniformly distributed, it can be shown that the distance of each of these points to any randomly chosen target point is expected to be approximately the same. If correlations and skew are added to the data, the sparsity level reduces, but not by a lot in many real settings. This means that the concept of spatial locality (as understood in the traditional definitions of distance) does not exist for high dimensional data. If such is the case, then how should nearest neighbors be defined? How should clusters be defined? How should outliers be defined? These high dimensional issues have been considered open questions. Furthermore, methods such as indexing are dependent upon meaningful spatial locality. When such spatial locality does not exist, the performance of index structures degrades rapidly.

5.2.1 In many cases, an alternative approach is to re-design the definition of distance based applications. This solves the dual problems of performance and meaningfulness simultaneously. For example, the clustering problem is re-defined as projected clustering a form of which is also known as local dimensionality reduction. The nearest neighbor search problem is re-defined as projected nearest neighbor search.

5.3 Privacy Preserving Data Mining:
The ubiquity and availability of personal information about users with many corporations has lead to concerns about the privacy aspects of data mining. An interesting area of research is to work with less information so as to preserve the privacy of individual users. Quantification of privacy preserving data mining algorithms are used. Methods for privacy-preserving data mining have also been extended to other domains such as text and binary data, and strings.

5.4 Data Stream Mining:
Recent advances in hardware technology have allowed us to store and record data at rates which were unimaginable a few years ago. For example, even a simple transaction such as using the credit card or the phone results in automated data storage. In many cases such as the traffic streams on the internet, one cannot even store the data, but can have small segments of the stream available for processing for a limited amount of time. In such cases, all data mining algorithms must satisfy a one-pass constraint. While the one-pass constraint is necessary, it does not fully capture many other unique aspects of data stream mining. For example, data streams show temporal locality in their distribution behavior. This temporal locality can be leveraged in order to improve the quality of the results.

5.5 Interactive Data Mining:
While computer science has come a long way in being able to perform a wide variety of automated tasks, there are limits to the abilities of a computer in some fields such as data mining. This is because many data mining algorithms do not have precise objective functions which can describe an application-driven task. Rather, the objective functions are artificially constructed in order to simulate our understanding of it. In many cases, the end results turn out to be surprising and quite unexpected. This is because an automated algorithm or mathematically defined objective function cannot simulate the intuition of the human mind. Clustering, classification, and nearest neighbor search are the human-computer co-operative algorithms. These methods combine the raw computational power of a computer with the intuition of the human effectively. The primary approach is to use visually driven methods to guide and support the user in making intelligent choices towards finding a solution for the data mining task at hand.

5.6 Web Crawling for Resource Discovery:
The web is a vast repository of documents and knowledge, waiting to be mined for a variety of purposes. The simplest application would be a search engine (such as Google or Yahoo!) in which crawlers regularly collect pages off the web and index them in order to facilitate user-centered keyword search. But what if you want an updated resource of web pages belonging to a particular topic? What if your query cannot be defined by the simple keyword interface of search engines? Recent work on web resource discovery has proposed methods for focussed crawling of topic specific documents. While focussed crawling is an interesting approach, it depends upon a fixed heuristic based on linkage structure. It is much more adaptive to use learning algorithms for focussed crawling and resource discovery. It helps in crawling the most appropriate web sites to crawl during the resource discovery process.

5.7 Data Mining for Electronic Commerce:
The ability to perform automated storage of web based transactions has resulted in vast repositories of useful data which can be mined for commercial purposes. Companies such as Amazon.com regularly use customer’s buying or browsing behavior in order to personalize their web site and portal presentations. Their aim is to maximize their revenue by presenting the advertisements and products which customer’s are likely to be interested in. This has given rise to the exciting field of data mining for electronic commerce. How do we use the previous customer browsing behavior to predict future behavior in a statistically robust way? Methods are used to perform WWW based advertising, Portal Personalization, and collaborative filtering.
5.8 Mining Specialized Data:

One of the interesting characteristics of the data mining area is that the field evolves very fast, since new data domains crop up frequently. This is because of new data collection and extraction technologies. For example, string mining has recently received a boost because of the ability to collect and extract genome data. Similarly XML processing has spawned the new area of structural mining. In some cases such as text mining, old domains become re-energized because of the effect of new enabling technologies such as the web. For each domain, the nature of the problems are quite different and require different algorithmic techniques. A number of algorithms for string classification, XML classification, indexing categorical data or market basket data have been developed.

6. Evaluating the Benefits of a Data Mining Model

Figure 1-1, which shows a "gains chart," suggests some benefits available through data mining. The diagonal line illustrates the number of responses expected from a randomly selected target audience. Under this scenario, the number of responses grows linearly with the target size.

The top curve represents the expected response. The target is now likely to include more positive responders than in a random selection of the same size. The shaded area between the curve and the line indicates the quality of the model. The steeper the curve, the better the model.

Other representations of the model often incorporate expected costs and expected revenues to provide the most important measure of model quality: profitability. A profitability graph such as Figure 1.2 can help determine the number of prospects to include in a campaign.

In this example, it is easy to see that contacting all customers will result in a net loss. However, selecting a threshold score of approximately 0.8 will maximize profitability.

7. CONCLUSION

Data Mining is the extraction of hidden predictive information from large databases. This is a new powerful technology with great potential to help companies focus on the most important information in data warehousing. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. "The automated, prospective analyses offered by data mining move beyond the analyzes of past events provided by retrospective tools typical of decision support systems."

Data mining is important to large systems because it finds things in large data repositories that users did not know existed.

Various challenging data mining and machine learning problems in advertising are addressed. A wide variety of topics are covered from pay-per-action business models for sponsored search, to effective targeting for content-based advertisement; from modeling user attention to classifying user sentiment, and from online advertisement to TV advertisement. The crucial issue of social challenges created by targeted advertisements is also addressed.

8. FUTURE SCOPE OF DATA MINING IN ADVERTISING

The future of data mining lies in predictive analytics. The technology innovations in data mining since 2000 have been truly Darwinian and show promise of consolidating and stabilizing around predictive analytics. Predictive analytics have successfully proliferated into applications to support customer recommendations, customer value and churn management, campaign optimization, and fraud detection. On
the product side, success stories in demand planning, just in time inventory and market basket optimization are a staple of predictive analytics. Predictive analytics should be used to get to know the customer, segment and predict customer behavior and forecast product demand and related market dynamics.

REFERENCES