Data Mining

Data Mining @ Google

Agenda of the Report

- **Abstract**
- **Introduction to data mining**
- **About Google- What it does**
- **Why Google uses data mining**
- **Technical details of implementation**
    - **Type of data used**
    - **Sourcing of data**
    - **Data Mining Tools used by Google**
- **Conclusion-Data Mining as a threat**

Abstract

This report introduces data mining to its audience by explaining data mining in the context of worlds’ most powerful search company- Google. Decades ago this area technique emerged as the most promising concept in the world of computer science and information technology. Since then dozens of search engines have been developed such as Yahoo, Google, Alta-Vista, Bing but the company which has been able to exploit the full power of this technology is Google whose search engine’s name was initially BackRub but was renamed to Google.

The main theme of this report is the usage of data mining by Google. Data mining is a process of discovering non-trivial information and patterns in a large collection of data. Emerging variants of data mining are able to perform multidimensional queries on a wide variety of heterogeneous data sources such as news websites, video search, image search or the historical manuscript search. The relevance of data mining in context of Google is the issue covered in this report.

Introduction to Data Mining

Data mining, a process better known as knowledge discovery, is the process of analyzing data from different perspectives and summarizing it into useful information-information which could be of economic value (which can increase revenue, cut costs or both). Companies use data mining as custom software by writing in-house code for analysis of large chunk of data. It can help us to identify patterns; it can categorize the data and find hidden relationships in it. It is a part of the process of knowledge discovery from relational databases. There has been a huge revolution in the IT industry in terms of information processing, it moved away from OLTP(Online transaction reporting) reporting to OLAP(Online Analytical Processing) reporting, OLAP was a very powerful technology and widely used by vendors and understood by developers. But there is so constant in innovative fields of computer science, the whole industry reached a stage where they needed another alternate of previous technologies, this was the dawn of data mining. Just like OLAP, it found its way from academic research to the real world where companies have invested billions to prepare their computing infrastructure to provide online services. Several tools offered these days on data mining are Oracle data mining, Weka, SQLServer data mining.

About Google-What it does

Google Inc. is a multinational company which started as an internet based search company, but soon expanded to wide range products company. They provide various internet based services such as online search, cloud computing based enterprise solutions, and advertising technologies. The company offers a free email solution-Gmail, an online office suite called Google docs, and they also offer online social networking solution names Google plus. Latest inception made by the company in browser software is Google Chrome, company has also developed an operating system called Chrome OS for netbooks and they also are the primary developer of open source mobile operating system called Android. The company is estimated to run around more than one million servers around the world and around one billion queries are made daily through its search service. Google also invests lot in development of new technologies apart from just using them and keeping this philosophy they have recently developed a new data mining tool called Correlate. This system offers companies and individuals to mine almost any kind of data leveraging the large number of algorithms developed by Google in previous years.

Why Google uses Data Mining

We have seen a complete revolution of data storage and retrieval in last two decades, where database technologies dominated the field. This search was revolutionized from simple text search which is stored in file to a search in hard disk, then text in relational databases, eventually leading to search of information of terabyte size in hard disks. But how does all that happen? How we get relevant and context aware information when we make a query to Google? Data mining answers this question. Google has developed in-house set of software tools and they don’t use only data mining for their information retrieval purpose, instead they use an integrated suite of tools and technologies which are given below:

- Artificial Intelligence
- Machine Learning
- Data Mining
- Neural Networks

- Knowledge recognition and acquisition
- Big Data
- Document based databases such as Mongo DB

All these technologies provide them huge power to run millions of queries every second which gives us information in the order of relevance to us.

They run a large cluster of machines which are responsible for archiving data using bots and crawlers. Archived pages are indexed by programs and then appropriate indexes are stored in databases which are finally searched by search system and information is provided to the users. The whole process may seem a little easy to a novice but internally it is a daunting process involving a whole set of services offered by Google. It is to be noted whenever we move today from one location to another location- the software helping us to navigate is GPS but the maps shown to us are provided by Google. In short the company is impacting almost everyone in the world in direct and indirect ways.

At Google, much of the companies' primary work relies on data mining whether it is basic information search, AdSense, ad words, maps, Google finance, their voice services, Gmail or enterprise based application suite provided by them- all exploit large scale data mining to a great extent. The requirement of data mining arises in various sorts of projects for example-machine based language translation, speech processing, and visual processing. This whole process involving machine learning at its core presents huge scientific and engineering challenges to them. The frequency of data collection is very large and the statistics change very rapidly which has to be updated in their databases in real time. Just like data collection, frequency of change in features of interest is quite large and the volume of data often precludes the use of standard single-machine training algorithms. The data mining systems placed at the core of rapidly expanding and highly dynamic services, they also use advanced statistical models which involve concepts from control systems and game theory also. The further portions of the report will discuss the implementation of data mining in Google.

Algorithms used by Google

Algorithms	**Characteristics**
Anomaly detection(One class support vector machine)	The model trains on data that is homogenous, that is all cases in one class, then determine if a new case is similar to the cases observed.E.g determining if a new case differs from known cases
Association rule(Apriori)	The model determines which cases are likely to be found

	together. E.g determining 5 items most likely to be purchased at the same time as item X(Market basket analysis)
Attribute importance(Minimum description length)	The model determines the importance of each attribute in predicting the target value. This is useful as a supplement to the main predicting model.
Classification(Naïve Bayes)	The model predicts two or more target values for each case.For e.g. Determining whether a customer will purchase a product X or whether a bank customer will default on a loan
Classification(Support vector machine)	Support vector machine(SVM) produces high quality result that is more complicated to use and takes longer to run
Clustering(K-means/ O-Cluster)	The model defines segments or clusters of a population, and then decides the likely cluster membership of each new case. E.g dividing year 12 exam results into bands and determining the attribute values indicating a student's falling into a particular brand.
Regression(Support vector machine)	The model predicts a specific target value for each case from among (possibly) infinitely many values. E.g determining that a product will be sold at a particular price.

The above list is not comprehensive but a small collection of algorithms used by Google for its data mining tasks.

Types Of Data Used and sourced By Google:

Google has, perhaps more than any other company, realized that information is power. Information about the Internet, information about innumerable trends, and information about its users, i.e us.

So how much does Google know about you and your online habits? It’s only when you sit down and actually start listing all of the various Google services you use on a regular basis that you begin to realize how much information you’re handing over to Google.

Now let’s have a look at how Google is gathering information from you, and about you.

Google’s information-gathering channels

Google’s stated mission is “to organize the world’s information and make it universally accessible and useful” and it is making good on this promise. However, Google is gathering even more information than most of us realize.

- **Searches (web, images, news, blogs, etc.)** – Google is, as you all know, the most popular search engine in the world with a market share of almost 70% (for example, 66% of searches in the US are made on Google). Google tracks all searches, and now with search becoming more and more personalized, this information is bound to grow increasingly detailed and user specific.
- **Clicks on search results** – Not only does Google get information on what we search for, it also gets to find out which search results we click on.
- **Web crawling** – Googlebot, Google’s web crawler, is a busy bee, continuously reading and indexing billions of web pages.
- **Website analytics** – Google Analytics is by far the most popular website analytics package out there. Due to being free and still supporting a number of advanced features, it’s used by a large percentage of the world’s websites.
- **Ad serving** – Adwords and Adsense are cornerstones of Google’s financial success, but they also provide Google with a lot of valuable data. Which ads are people clicking on, which keywords are advertisers bidding on, and which ones are worth the most? All of this is useful information.
- **Email** – Gmail is one of the three largest email services in the world, together with competing options from Microsoft (Hotmail) and Yahoo. Email content, both sent and received, is parsed and analyzed. Even from a security standpoint this is a great service for Google. Google’s email security service, Postini, gets a huge amount of data about spam, malware and email security trends from the huge mass of Gmail users.
- **Twitter** – Google has direct access to all tweets that pass through Twitter after a deal made late last year.
- **Google Apps (Docs, Spreadsheets, Calendar, etc.)** – Google’s office suite has many users and is of course a valuable data source to Google.

- **Google Public Profiles** – Google encourages you to put a profile about yourself publicly on the Web, including where you can be found on social media sites and your homepage, etc.
- **Orkut** – Google's social network isn't a success everywhere, but it's huge in some parts of the world (mainly Brazil and India).
- **Google Public DNS** – Google's newly launched DNS service doesn't just help people get fast DNS lookups, it helps Google too, because it will get a ton of statistics from this, for example what websites people access.
- **The Google Chrome browser** – What is your web browsing behavior? What sites do you visit?
- **Google Finance** – Aside from the finance data itself, what users search for and use on Google Finance is sure to be valuable data to Google.
- **YouTube** – The world's largest and most popular video site by far is, as you know, owned by Google. It gives Google a huge amount of information about its users' viewing habits.
- **Google Translate** – Helps Google perfect its natural language parsing and translation.
- **Google Books** – Not huge for now, but has the potential to help Google figure out what people are reading and want to read.
- **Google Reader** – By far the most popular feed reader in the world. What RSS feeds do you subscribe to? What blog posts do you read? Google will know.
- **Feedburner** – Most blogs use Feedburner to publicize their RSS feeds, and every Feedburner link is tracked by Google.
- **Google Maps and Google Earth** – What parts of the world are you interested in?
- **Your contact network** – Your contacts in Google Talk, Gmail, etc, make up an intricate network of users. And if those also use Google, the network can be mapped even further. We don't know if Google does this, but the data is there for the taking.
- **Coming soon** – Chrome OS, Google Wave, more up-and-coming products from Google.

Much of this data is anonymized, but not always right away. Logs are kept for nine months, and cookies (for services that use them) aren't anonymized until after 18 months. Even after that, the sheer amount of generic user data that Google has on its hands is a huge competitive advantage against most other companies, a perpetual gold mine.

Google-Data Mining Tools:

Google correlate

Google Correlate is like Google Trends in reverse. With Google Trends, you type in a query and get back a data series of activity (over time or in each US state). With Google Correlate, you enter a data series (the target) and get back a list of queries whose data series follows a similar pattern.

The objective of Google Correlate is to surface the queries in the database whose spatial or temporal pattern is most highly correlated (R square) with a target pattern. Google Correlate employs a novel approximate nearest neighbor (ANN) algorithm over millions of candidate queries in an online search tree to produce results similar to the batch-based approach employed by Google Flu Trends but in a fraction of a second.

Approximate Nearest Neighbor (ANN) system achieves a good balance of precision and speed by using a two-pass hash-based system. In the first pass, it computes an approximate distance from the target series to a hash of each series in database. In the second pass, it computes the exact distance function on the top results returned from the first pass. Each query is described as a series in a high-dimensional space. Since the number of queries in the database is in the tens of millions, computing the exact correlation between the target series and each database series is costly. To make search feasible at a large scale, ANN system is employed that allows fast and efficient search in high-dimensional spaces.

Traditional tree-based nearest neighbors search methods are not appropriate for Google Correlate due to the high dimensionality which results in sparseness. Most of these methods reduce to brute force linear search with such data. For Google Correlate, novel asymmetric hashing technique is used which uses the concept of projected quantization to reduce the search complexity. The core idea behind projected quantization is to exploit the clustered nature of the data, typically observed with various real-world applications. At the training time, the database query series are projected in to a set of lower dimensional spaces. Each set of projections is further quantized using a clustering method such as K-means. K-means is appropriate when the distance between two series is given by Euclidean distance. Since Pearson correlation can be easily converted into Euclidean distance by normalizing each series to be a standard Gaussian (mean of zero, variance of one) followed by a simple scaling, K-means clustering gives good quantization performance with the Google Correlate data. Next, each series in the database is represented by the center of the corresponding cluster.

Filtering

Google Correlate makes an attempt to filter out queries which are unlikely to be interesting. These include:

- Queries with a low correlation value (less than *r*=0.6)
- Misspelt queries
- Adult queries
- Rare queries

- Queries which only correlate with a small portion of the time series

Data mining as threat

An imminent report on an emerging threat to individual privacy to be issued by the European data protection authorities raises even more serious issues than those it is likely to address. The report will consider Google's asserted right to expand its data mining to combine users' personal data across all their accounts and services, including Gmail, internet searching, map and location information and photo sharing, with no way for individuals to opt out. At least one technology blogger has accused Microsoft of planning similar changes, while two new Facebook programmes to aggregate user data with other advertising and loyalty card data have also drawn concern. Whatever the merits of each case, the larger issue deserves greater public attention.

There is a powerful reason why cloud services and other data-mining companies aggregate data across multiple accounts and services: the results are extremely valuable. Just as tiny bits of coloured tile can be combined and transformed into a coherent piece of art, tiny bits of seemingly unrelated personal data, when aggregated and mined at huge scale, can provide immense value to advertisers, marketers, corporate sales forces and others. The revenue generated by combining and monetising such data – by mining the mosaic – is the reason "free" cloud services can afford to be free.

Privacy groups and regulators are appropriately concerned with threats to individual privacy inherent in mosaic-mining business models. Less noticed has been the potential use of these same tools and techniques against government employees and, potentially, governments themselves. But is this a more serious problem and, if so, why? The privacy rights of government employees are no more or less important than those of private citizens. Beyond individual privacy, however, consider the national security, government integrity, and even personal safety implications.

What if, instead of using the power of mosaic mining to identify a potential new customer, it was used to identify an undercover intelligence operative? Multiple map queries starting at a known intelligence agency's headquarters instead of the work address on an operative's business card, might suggest their true job, even if those queries were conducted from their personal computer. Such a scenario is neither hypothetical nor attractive just to adversarial intelligence services. American lawyers defending Guantánamo inmates provided surreptitiously obtained photographs of undercover CIA operatives to their clients. Whatever the lawyers' motives, imagine how much more could be done by those motivated to out covert agents, if they mined today's mosaic of private and governmental information.

What if, instead of mining the mosaic to anticipate a company's office supply needs, data from across individual and government accounts – email contents, internet searches, travel plans – were used to anticipate a government policy decision or treaty negotiating position? What might the geolocation of a cop's personal laptop for several consecutive nights reveal about a stakeout? And what pressure could be put on government decision-makers by someone armed with video download receipts, browser search records or credit card statements?

These risks do not assume ill motives on the part of cloud service providers, although we must assume that there are at least a few such companies around the globe not particularly diligent about selling amalgamated customer data. But no company can hope to block all ill-motivated insiders or skilled hackers. Massive databases of aggregated personal and governmental data would present irresistible targets.

Beyond such security concerns, what European data protection authorities are confronting is an early, but vital, test of whether governments will continue to control their own data. Governments considering deployment of cloud computing solutions should consider several steps to mitigate these risks. The first step is awareness. Law enforcement, security and other governmental organizations should consider carefully the ramifications of a private entity being able to aggregate sensitive data across government, and their employees', private accounts. Second, governmental entities should not accept generic provider privacy policies, but should demand government-specific agreements prohibiting data mining.

Governments also should insist that data-mining capabilities be technologically disabled from use against their data. Providers whose business models are so dependent on data mining that they cannot compete for government business without it may have to stick to the many lucrative non-governmental markets. Beyond this, governments should ensure they can determine independently whether vendors are living up to no-data-mining assurances. Finally, governments should provide their employees with awareness training about the risks of data mining of their personal and governmental accounts – a prudent measure whether or not cloud solutions are deployed.

Personal privacy is vitally important. But if governments do not address the national security implications of mining the mosaic, we may soon look back with nostalgia to when personal privacy was our only concern.