Big Data and Predictive Healthcare

Vaikkunth Mugunthan

Abstract - We live in an on-demand, on-command Digital universe with data proliferating by Institutions, Individuals and Machines at a very high rate. This data is categorized as "Big Data" due to its sheer Volume, Variety, Velocity and Veracity. Most of this data is unstructured, quasi structured or semi structured and it is heterogeneous in nature. Human body do generate lot of data which can be captured at regular intervals and when combined with other medical records (EHR) contribute to large volume and variety of data (big data).

On one side , technology continues to evolve, population continues to grow, medical facilities increase day-by-day and lot of information available. However, on the other side, this useful information cannot be made use of to a greater extent as information is scattered, unorganized, analyzing of the same is complex an above all the proportion of doctors/specialists to patients is very low.

Further, issues like obesity has been found to be very common in US and other countries, predominantly in children below age 14 which could lead to major health concerns at later ages.

In this paper, we will have a look at some of the ways to capture the data, relevant technologies involved and some ways of doing the predictive analytics towards proactive health monitoring and preventing diseases.

Index Terms— Big data, Euclidean distance, faceted search, IOT, kMSP, predictive analytics, Solr , Sphinx ,wearable devices

---

INTRODUCTION - "Big Data" is typically considered to be a data collection that has grown so large it can't be effectively or affordably managed (or exploited) using conventional data management tools: e.g., classic relational database management systems (RDBMS) or conventional search engines, depending on the task at hand. This can as easily occur at 1 terabyte per hour to exabytes based on the data collection interval and the nature of business.

"Big Data" is really a specially coined term to convey the extraordinary scale of the data collections now being amassed inside public and private organizations and out on the Web.

Familiar Challenges, New Opportunities:

Big Data is not new for information specialists and analysts in fields like banking, telecommunications and the physical sciences. They have been grappling with Big Data for decades and have confronted data collections that outgrew the capacity of their existing systems. In most of the situations their choices were always less than ideal:

- Need to access it? Segment (silo) it.
- Need to process it? Buy a supercomputer.
- Need to analyze it? Will a sample set do?
- Want to store it? Forget it: use, purge, and move on.

Vaikkunth Mugunthan is currently pursuing Bachelor's degree program in Information Technology in SSN College of Engineering, India, PH-8939370191. E-mail: krishna051295@gmail.com

natural information intelligence locked inside our

New technologies have emerged which allows organizations of any type to analyze and exploit Bigdata which may include data that was too voluminous, complex or fast-moving to be of much use before (e.g meter or sensor readings from IOT or wearable devices, event logs, Web pages, social network content, email messages and multimedia files).

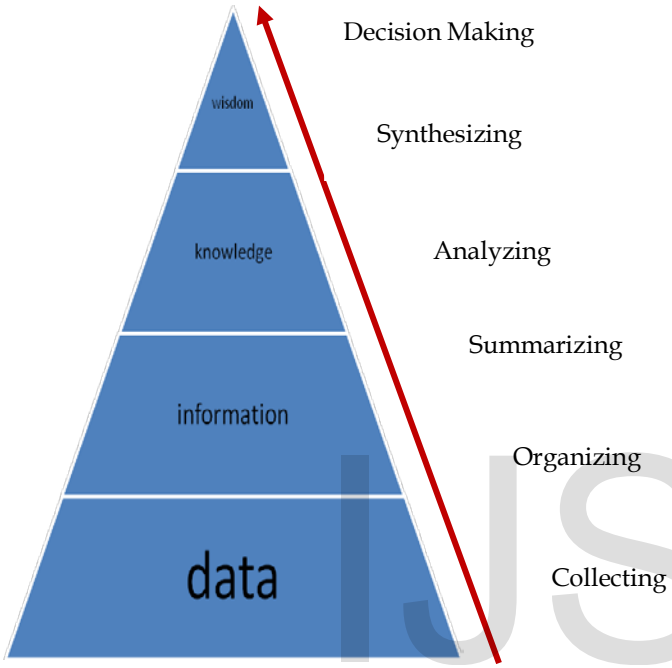
As a result of this evolution, the Big Data universe is beginning to yield insights that are changing the way we work and the way we play, and challenging just about everything we thought we knew about ourselves, the organizations in which we work, the markets in which we operate – even the universe in which we live.

Big Data – Affected entities : Big Data has been of concern to organizations working in select fields for some time, such as the physical sciences (meteorology, physics), life sciences (genomics, biomedical research), government (defense, treasury), finance and banking (transaction processing, trade analytics), communications (call records, network traffic data), and, of course, the Internet (search engine indexation, social networks).

Big Data: Boon or Bane? For some organizations, Big Data simply means Big Headaches, raising difficult issues of information system cost (Total cost of ownership), scalability and performance, as well as data security, privacy and ownership. However, breakthrough insights and innovation in business, science, medicine and government can occur with BigData analytics when we can bring humans, machines and data together to reveal the mountains of Big Data.

Steps involved: The following steps pertain to transforming raw data into action-guiding wisdom:

Decision Making
Synthesizing
Analyzing
Summarizing
Organizing
Collecting



Data Management steps

Big Data Opportunities: Many organizations are opening new frontiers in operational and exploratory analytics using structured data (like database content), semi-structured data (such as log files or XML files) and unstructured content (like text documents or Web pages). Some of the specific Big Data opportunities they are capitalizing on include:

- Faceted search at scale
- Multimedia search
- Sentiment analysis
- Automatic database enrichment
- New types of exploratory analytics
- Improved operational reporting

Faceted Search at Scale

Faceted search is the process of iteratively refining a everything from ideas and issues to people, products and companies.

search request by selecting (or excluding) clusters or categories of results. In contrast to the conventional method of paging through simple lists of results, faceted search (also referred to as parametric search and faceted navigation) offers a remarkably effective means of searching and navigating large volumes of information—especially when combined with user aids like type-ahead query suggestions, auto-spelling correction and fuzzy matching (matching via synonyms, phonetics and approximate spelling).

Until recently, faceted search could only be provided against relatively small data sets because the data classification and descriptive meta-tagging upon which faceted search depends were largely manual processes. Now, however, industrial-grade natural language processing (NLP) technologies are making it possible to automatically classify and categorize even Big Data-size collections of unstructured content, and hence to achieve faceted search at scale.

Public Web search engines like Google, Yahoo! and Bing, and, to varying degrees of automation and scale, in search utilities from organizations like HP, Oracle, Microsoft and Apache. This trend will continue to accelerate to bring new accessibility to unstructured Big Data.

Multimedia Search: Multimedia content is the fastest growing type of user-generated content, with millions of photos, audio files and videos uploaded to the Web and enterprise servers daily. Exploiting this type of content at Big Data scale is impossible if we must rely solely on human tagging or basic associated metadata like file names to access and understand content.

However, recent technologies like automatic speech-to-text transcription and object-recognition processing (called Content-Based Image Retrieval or CBIR) are enabling us to structure this content from the inside out, and paving the way toward new accessibility for large-volume multimedia collections. This trend will have a significant impact in fields like medicine, media, publishing, environmental science, forensics and digital asset management.

Sentiment Analysis: Sentiment analysis uses semantic technologies to automatically discover, extract and summarize the emotions and attitudes expressed in unstructured content. Semantic analysis is sometimes applied to behind-the-firewall content like email messages, call recordings and customer/constituent surveys. More commonly, however, it is applied to the Web, the world's first and foremost Big Data collection and the most comprehensive repository of public sentiment concerning

Sentiment analysis on the Web typically entails

collecting data from select Web sources (industry sites, the media, blogs, forums, social networks, etc.), cross-referencing this content with target entities represented in internal systems (services, products, people, programs, etc.), and extracting and summarizing the sentiments expressed in this cross-referenced content.

Database Enrichment: Once you can collect, analyze and organize unstructured BigData, you can use it to enhance and contextualize existing structured data resources like databases and data warehouses.

For instance, one can use information extracted from high-volume sources like email, chat, website logs and social networks to enrich customer profiles in a Customer Relationship Management (CRM) system. Or, can extend a digital product catalog with Web content (like, product descriptions, photos, specifications, and supplier information). Such content can also be used to improve the quality of the organization's master data management, using the Web to verify details or fill in missing attributes.

Exploratory Analytics: Exploratory analytics has aptly been defined as "the process of analyzing data to learn about what you don't know to ask." It is a type of analytics that requires an open mind and a healthy sense of curiosity. In practice, the analyst and the data engage in a two-way conversation, with researchers making discoveries and uncovering possibilities as they follow their curiosity from one intriguing fact to another (hence the reason exploratory analytics are also called "iterative analytics").

In short, it is the opposite of conventional analytics, referred to as Online Analytical Processing (OLAP). In classic OLAP, one seeks to retrieve answers to precise, pre-formulated questions from an orderly, well-known universe of data. Classic OLAP is also sometimes referred to as Confirmatory Data Analysis

Operational Analytics: While exploratory analytics are terrific for planning, operational analytics are ideal for action. The goal of such analytics is to deliver actionable intelligence on meaningful operational metrics in real or near-real time. This is not easy as many such metrics are embedded in massive streams of small-packet data produced by networked devices like 'smart' utility meters, RFID readers, barcode scanners, website activity monitors and GPS tracking units. It is machine data designed for use by other machines, not humans. Making it accessible to human beings has traditionally not been technically or economically feasible for many organizations. New technologies, however, are enabling organizations to overcome technical and financial hurdles to deliver human-friendly and analysis of real-time Big Data streams. More organizations (particularly in sectors like telecom, transport, retail and manufacturing) are producing real-time operational reporting and analytics based on such data, and significantly improving agility, operational visibility, and day-to-day decision making.

Tools primarily used for BigData Analysis:

Some of the most popular tools/technologies used for BigData analysis include the following:

NoSQL (Databases - MongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper)

MapReduce (Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum)

Storage (S3, Hadoop Distributed File System)

Servers (EC2, Google App Engine, Elastic, Beanstalk, Heroku)

Processing (R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop)

Business benefits: A telecommunications company wants to reduce its risk of customer loss, so it analyzes billions of call detail records to find out which customers are the most connected (that is, make or receive the most calls from a wide variety of phone numbers). The company then focuses promotions on these individuals to keep them as happy customers, since if they leave, they may "drag" a lot of friends with them to a new carrier. It is this type of hidden insight that demonstrates how big data expands the range of information used in decision making. Enterprises can now create new business value by leveraging sources of data that were previously hard to capture, access and analyze because of challenges with its size, speed and structure.

Three Categories of Business Opportunities

Big data can unlock new business value in a wide variety of ways, but most notably in three types of opportunities: Making better-informed decisions, discovering hidden insights and automating business processes.

Better-Informed Decisions: In the first case, decisions such as prices, promotions, staffing levels or investments — any business decision — can be improved if big data sources are available for insight. Take for example, Wal-Mart, which wanted to help its website shoppers find what they were looking for more quickly. It developed a machine learning semantic search capability using clickstream data from its 45 million monthly online shoppers combined with product- and category-related popularity scores generated from text mining social media streams. Wal-Mart's resultant Polaris" search engine yielded a 10% to 15% increase in online shoppers completing a purchase (or around a billion dollars in incremental sales).

Hidden Insights: Big data analysis can also be used to discover opportunities that are obvious only by looking at large sets of detailed data. Many organizations are mining vast pools of data to discover hidden insights that were previously unavailable to them – often in the development of new or enhanced products.

The Climate Corp was started by former Google employees to offer crop insurance to underserved parts of the world. It continually gathers weather and soil measurements from 500,000 locations and has 30 trillion data points to date. Complex analytics predict weather related risks for specific crops in specific locations. This enables Climate Corp to outcompete other insurers that cannot assess risk at that level of locale specificity, and enables farmers in Asia and Africa to take on the risk of buying seeds, labor and farm equipment that they otherwise could not.

Automate Business Processes: Finally, new technology can be used to leverage big data in real time, allowing analysis to be built into processes so that automated decision making can occur. One of McDonald's bakeries replaced calipers and color cards with high-speed image analytics to scrutinize thousands of buns per minute for color, size and even sesame seed distribution — instantly adjusting oven and other process controls to create uniform buns and reduce wastage. Another food products company similarly photo analyzes and sorts each and every French fry produced to optimize quality.

Tapping into Dark Data: An enterprise that is proficient in analyzing big data has a whole new world of data sources available. It can now leverage data both inside and outside the enterprise that was previously unavailable or not utilized. For example, inside the enterprise, there exist underutilized datasets, or "dark data." These can include email archives, warranty forms, call center recordings and doctors' notes. Large public sources of data such as social media and government also become a source of potential value.

Risk Management: Big data is very often used for fraud management. Graph analytics, in particular, can help in detecting fraud rings. By understanding relationships in the data, hidden commonalities can be discovered. In some instances, people who commit fraud share a certain location, are from a similar age group, or are seemingly disconnected companies sharing common ownership. The results are real. Not only in terms of business value, but they also lead to societal impact. The newspapers recently reported that in Europe, where healthcare is often less privatized, various insurance companies have uncovered claims fraud by dentists and other healthcare providers who were using diagnostic codes that didn't match real procedures.

Banks often use big data for improving credit scoring, using graph analytics to include social relationships as an indicator of risk, or better yet, credibility. In an ironic twist, modern technology is reinventing old values in community banking to provide knowledge that a certain family is good for its money, even though family members growing up haven't shown that behavior themselves yet.

New Business: Perhaps the most exciting business cases come from a new value discipline: treating information as a product in itself. Utilities and banks already provide their customers with personalized dashboards about their use of financial products or energy. Remote patient monitoring is a growth business for healthcare providers or life science companies. Wearable computing introduced a category called "personal analytics," where consumers can measure and share health indicators such as heart rate, blood pressure and calorie consumption.

Industry Inspiration: Success falls to those organizations who creatively embrace big data. However, it's also important to note that inspiration often comes from other industries. A big data scenario from travel and transportation can, for instance, be used in retail for automated store task management.

In this instance, a train company uses video feeds of its cabin safety cameras to provide information to travelers at the next station about which cabins still have enough seats available. A retailer can leverage this to count how many people enter the supermarket, combine that data with an

understanding of the average shopping time on a weekend in rainy weather conditions to help predict when to open up a new cash register.

In another example, malls can learn from the games industry. By tracking through a smartphone where mall visitors are, coupons can be shared on the spot. Additionally, visitors earn points by tracking how many stores they visit and are promoted to the next level.

Analytics Will Enable Better Decision Making

Decisions are a basic unit of work for all organizations. The success of every enterprise is a function of the cumulative effect of the quality of the decisions that it makes. Despite large BI investments in the name of better decision making, poor decisions are abound. Where decision rules and logic are well known, more precise and real time analytics will be applied to automate a range of operational decisions.

For example, a retail food chain monitors refrigeration assets in real time to proactively predict and maintain an asset before it fails. At the same time, the quality of collaborative decisions and professional experiential and judgment-based decisions (clinical diagnosis, employee hiring, online education, personal health and wellness) will be enhanced by advanced analytics, man-machine partnerships or digital assistant models (think IBM Watson); and many more are emerging.

kMSP - k-means for Most Suitable Product (Profile) algorithm

Now a days, most of the ecommerce providers like Amazon, Flipkart etc., make use of the "recommendation systems" to suggest the relevant products based on the users' previous purchase, searches made etc.,

I am currently working upon an algorithm similar to recommendation systems with a different approach to the search algorithm based on "scores/weights" for various attributes.

This algorithm is a kind of modification of the "nearest neighbor algorithm. Here, we assign weights/score to each of the attributes and based on the search attributes/keywords, user's preferences, attributes of the profiles/products. The algorithm will calculate the final rank/score of the matching profiles/products and render the search results, beginning with the "closest matching" values at the top. An ideal profile/product will have a score of "zero" i.e. closer to zero is "highly matching".

The algorithm makes use of Euclidean distance formula in combination with nearest neighbor algorithm. The distance (closeness) can be computed like below:

For each of the profiles/products in the database, we calculate the distance score (d) of the profile/product from the input entered by the active user (who is making the search) , by using the Euclidean Distance formula. The ED formula is given below:

d (x,y) = square root of (w1(x1-y1)^2+w2(x2-y2)^2+….+wn(xn-yn)^2)

The values of x1, x2, x3…xn are all set to '10'. The value of "n" can be set as desired (i.e the number of attributes to be matched).

As an example, x1 , x2 and x3 could represent age, height and weight respectively of the similar/matching profiles (based on a 10 point scale system) and y1,y2 and y3 would represent similar values of the active user while w1,w2 etc. could represent the relevant weightage for the attributes.

The algorithm will be beneficial in obtaining very closely matching profiles/products/medicines based on the search criteria. Tools and technologies like Solr / Sphinx can be used to render the search results with Mysql as a possible database and these are completely open-source.

Application of the algorithm: This algorithm can potentially be used in the area of healthcare, dating, selection of products etc., We can easily identify the right medicine for a patient by matching his/her attributes with the matching attributes of an individual whose profile is similar by searching the database and obtain the relevant medicine.

For example, if we have to find out the right medicine for a patient by passing attributes like "citizenship, gender, age, habits like smoking/non-smoking, food type like vegetarian/non-vegetarian, blood group etc.," and search the database to identify the relevant list of illnesses that he/she may potentially get and/or the right medications (if the illness can also be passed as a search criteria).

Along the same concept, suitable partner profiles or matching products to the "maximum fit" can be identified easily related to dating or shopping respectively.

As the database size increases and by suitably modifying the "training set" i.e. machine learning could be of great help in coming out with the "best fit" more easily.

CONCLUSION: In general, there are three ways to use big data visualization as part of any organization's

corporate strategy: as a core product or service offering (e.g., Cloudera), as a supporting product or service (e.g., Progressive), or as a key organizational capability. If one is employing one of the first two approaches, the use of big data analytics should be conspicuous. If one is trying to build a big-data-driven organization, that message may not be so obvious. In this case, organizational change management is extremely important to your strategy, and it begins with a very basic step that's often taken for granted: awareness.

In conclusion, big data is not about just handling volume, nor is it about data. It is about creativity. Combine technology advancements with human ingenuity and the possibilities are endless. While there are various tools and technologies available for processing BigData, it all boils down to the core objective(s) or the strategy of any organization that might decide the results as "Boon or Bane", considering the time, efforts and the cost involved in all stages of the analysis.

REFERENCES

[1] Scott Spangler, IBM Almaden Services Research, "A Smarter Process for Sensing the Information Space," October 2010
[2] http://en.wikipedia.org/wiki/Bigdata
[3] Constance Hays, "What Wal-Mart Knows About Customers' Habits," The New York Times, November 14, 2004.
[4] Book - "Search-Based Applications: At the Confluence of Search and Database Technologies," Gregory Grefenstette and Laura Wilber,
Morgan & Claypool Publishers, December 2010.
[5] www.nosqldatabase.org
[6] www.vldb.org.
[7] Various research papers of Gartner Analysts and Articles from Harvard Business Review
[8] Whitepapers, tutorials on Big Data of various contributors on internet