INTERESTED IN... BIG DATA
SIMULATION OPTIMISATION
RISK ANALYSIS VISUALISATION
PREDICTIVE MODELLING ADVANCED ANALYTICS
DATA MINING DATA SCIENCE
CUSTOMER INSIGHT FORECASTING
...this is the professional magazine for you

AQ
Analytics Quarterly
Winter 2013

THE OR SOCIETY

the Analytics Network
a section of the OR Society

Is predicting the future, fantasy or good mathematics?

Nigel Cummings

The Arab Spring, a term given to the Arab Revolution was a revolutionary wave of demonstrations and protests which occurred in the Arab world from 18 December 2010 but was it predictable?

To date, rulers have been forced from power in Tunisia, Egypt, Libya, and Yemen; and probably very soon in Syria. Additionally civil uprisings have erupted in Bahrain and major protests have broken out in Algeria, Iraq, Jordan, Kuwait, Morocco, and Sudan. Yet for some reason intelligence officials using the latest predictive analytics applications, failed to predict any of these uprisings. All this has occurred despite the fact that billions of pieces of disparate electronic information have been analysed in order to identify hot spots before they explode.

The intelligence community has always been in the business of forecasting the future. The question now though, is whether tapping into publicly

available data such as twitter, news feeds, social networks and blogs can help them do that faster and more precisely. Surely the information dredged from such sources could have been used to identify potential hotspots developing in the Arab world?

Perhaps the problem with predicting such events has been in the type and quality of analytics used to analyse the data. Now in what seems to be echoes of the popular science fiction film, 'Minority Report' a Swedish-American start-up company called Recorded Future has developed algorithms that sift through huge volumes of information to find relationships between people and organisations. Then the sifted and sorted data is processed by 'visualisation software' to generate giant searchable timelines of possible future events.

Co-founder of Recorded Future, Christopher Ahlberg, a former member of the Swedish Special Forces, says. "What we're trying to do here is figure out a cool way that we can use to observe the world. We're trying to find new ways of generating data that tell us what's going on in the world ... what did happen, what will happen. We're not going to get 100% in terms of outcome, but we can pull things together in a way that no one else can. So, what we are trying to do is figure out how we can take large portions of the web and extract what we call signals of activity that relate to people and places and associate them with events and time. Time is often a forgotten dimension in analysis, and we think it is key."

Recorded Future's goal is to provide analytical tools which give users the ability to see events or relationships in sequence and make it easier to find patterns and relationships that traditional 'Big Data' programs might miss. Hedge funds already use Recorded Future to invest, and the intelligence community could use it to help predict world events.

Ahlberg says there are hints about the future everywhere. Governments release economic projections; newspapers report on upcoming events; and information derived from Twitter can provide a good idea of what people are talking about. In Egypt last year, organisers used Twitter and social media to rally protestors. If intelligence analysts had had a systematic way to track those posts, it might have helped them forecast what was to come.

There have already been efforts to try to tap into what is bubbling under the surface by tracking things like Google searches, and Researchers at West Point's Combatting Terrorism Centre and Princeton University tracked Google searches in Egypt starting in January 2011 and found, for example, that there were more searches about events in Tunisia and its protests than for Egyptian pop stars. Recorded Future builds on that kind of public intelligence.

Utilising such data and adding the element of time could be the key to gaining useful predictions from such data. Ahlberg also says that. "An event we'd track might be people traveling from A to B ... people talking to each other ... a government guy making a statement, a country doing a military manoeuvre. We capture those activities ... they can be small scale or large scale - and then time is associated with that. What we're doing is organising the data in a way so you can ask the right questions of it."

In January 2010, Recorded Future predicted from a number of blog posts that Yemen was headed for disaster. It predicted that a combination of floods, famine and Islamic terrorists were conspiring to wreak havoc. The company used Twitter feeds, blogs, U.N. food program data and news sources to come up with their forecast. According to Ahlberg, accuracy of that particular time line was not 100% though as, "Yemen took five months longer than we predicted, but if you go back and look at our earliest blog posts, it's all there. That said, we didn't predict the Arab Spring."

After the attack on the U.S. consulate in Libya, Recorded Future made a prediction that the attack would eventually be traced to al-Qaida's arm in Yemen, known as al-Qaida in the Arabian Peninsula. That is the same al-Qaida group that was behind the Christmas Day attack on a U.S. airliner in 2009. Recorded Future came to that conclusion because its algorithm linked a militant group in Libya, Ansar al-Sharia, with an al-Qaida group in Yemen with the same name. The people who use Recorded Future, Ahlberg says, are experts, so they should pick this up.

Of course, now that people have been made aware of this type of analysis, they may use it to their advantage by generating large amounts of false information to mislead the analysts as to when and where the next terrorist activity will take place in much the same way as Operation Bodyguard in WWII.

AQ

You're so 'data scientist' and you don't even know it!

Nigel Cummings

Thanks to Sue Merchant, (Blue Link Consulting), for bringing this material to my attention.

Increasing numbers of us are signing up to LinkedIn the world's largest professional network with over 175 million members and growing rapidly. LinkedIn connects us to trusted contacts and helps in the exchange of knowledge, ideas, and opportunities amongst a broad network of professionals. In June 2006 when LinkedIn was in its infancy though, it fell to Jonathan Goldman, PhD in physics from Stanford, as if he were arriving for work in a business that was still in its start-up phase.

At that time LinkedIn had less than 8 million accounts, but it was clear there was some growth in the signing up of new members as existing members were inviting their friends and colleagues to join. Unfortunately site users were not seeking out connections amongst existing LinkedIn members; they were merely assisting in the creation of new members.

A LinkedIn manager at that time, summed up what was missing from the site's operation, "It was like arriving at a conference reception and realising you don't know anyone. So you just stand in the corner sipping your drink -and you probably leave early".

That statement probably equates well to those who join social networking sites such as LinkedIn, Facebook, or Workday because of an invite, but fail to stay connected because they fail to see the point of belonging to such networks.

Goldman decided to look into the phenomena that were increasing membership but not usage because he was intrigued by the linking he could see taking place and by the richness of the user profiles. From his observations it became clear to him that allowing existing members to carry on in in this way was a recipe for messy data and in the long term, for unwieldy analysis. So he began exploring LinkedIn people's connections and started to see some possibilities where order could prevail.

He began forming theories, and found patterns that allowed him to predict whose networks a given profile would land in - he could imagine that new features capitalising on the heuristics he was developing might provide value to users. Unfortunately, at that time, LinkedIn's engineering team were more concerned with scaling up the site, than exploring Goldman's ideas. Some of the team wondered why users would ever need LinkedIn to figure out their networks for them. The site, after all, had an address book importer that could pull in all a member's connections.

Fortunately LinkedIn's co-founder Reid Hoffman, LinkedIn's cofounder, had faith in the power of analytics and because of his earlier experiences with analytics whilst working at PayPal, he decided to give Goldman a high degree of autonomy which would allow him to circumvent the companies' traditional product release cycle and allow him to publish small modules in the form of ads on the site's most popular pages.

One of Goldman's early modules looked at what would happen if users were presented with names

of people they hadn't yet connected with but seemed likely to know - people who had shared their tenures at schools and workplaces for example. The module he selected to test his theories displayed the three best new matches for each user based on the background entered in his or her LinkedIn profile. Within a short time of implementing his module – the click-through rate on each module became the highest ever seen for LinkedIn.

Goldman continued to refine how the suggestions in his modules were generated, incorporating networking ideas such as "triangle closing". The idea there, that should you know two members for example, there would be a good chance that those two members would know each other too.

This idea was so good that LinkedIn made it a standard feature before long, and "People You May Know" ads achieved a click-through rate 30% higher than the rate obtained by other prompts to visit more pages on the site. This in turn generated millions of new page views and caused LinkedIn to enlarge its growth curve upward – the start-up had it seemed, come of age!

Jonathan Goldman is an example of a new type of data professional; he is a "data scientist", a high-ranking professional with the experience and curiosity to make discoveries in the world of big data. The term data scientist by the way is a relatively new it was coined sometime during 2008 by D.J. Patil, and Jeff Hammerbacher, then the respective leads of data and analytics efforts at LinkedIn and Facebook.

Without knowing it though, thousands of us are working as 'data scientists' at both start-ups and well-established companies. Our sudden 'importance' to the business scene reflects the fact that companies are having to come to terms with increasingly large amounts and varieties of information – better known perhaps as Big Data!

Much of the current enthusiasm for big data focuses on technologies that make taming it possible, including Hadoop (the most widely used framework for distributed file system processing) and related open-source tools, cloud computing, and data visualisation. Not only but also, here is the good news there is a shortage of data scientists in many business sectors and O.R. professionals are natural born data scientists, we've been doing it for years!

Companies are increasingly looking to recruit data scientists. If capitalising on big data depends on hiring scarce data scientists, then the challenge for managers is to learn how to identify that talent, attract it to their enterprise, and make it productive. The need for growing numbers of such

Much of the current enthusiasm for big data focuses on technologies that make taming it possible

scientists could be good news for O.R. recruitment in 2013. We are natural born problem solvers with natural born data processing skills – what a potent combination?

Interestingly, there are no university programs offering degrees in data science yet unless of course you consider some of the components and modules that O.R. students encounter in their university studies. There is also little consensus on where the role of data scientists fit into organisations yet, and how data scientists can add value. These could be golden times for O.R. professionals to augment their experience portfolios and prove to the outside world, the increasing value of analytics and associated O.R. skills.

AQ

What's next for Analytics?

Nigel Cummings

Precision, personalisation and collaboration are on the agenda for the next wave of applied analytics.

Research done by Gartner suggests that data warehouse layers and OLAP cubes will still have a place with enterprise data, though precision, personalisation and collaboration are likely to be the prime focus in the next generation of analytics applications.

According to Kurt Schlegel, analytics research VP for Gartner, for more precise analytics, data architects should look to augment rather than replace existing operations with an eye toward agility and acceptance of external and varied data sources.

There will be, he said, more acceptance of the use of data discovery tools, especially where there is little training required and as big data frameworks become more visually appealing. This provides an opportunity for industry-specific, trusted data aggregators to fill in enterprise data gaps with as-a-service options.

Already, existing providers of benchmarking and industry analysis are putting together niche offerings that provide the content that many enterprises may not be able to for lack of in-house talent, budget or existing data resources. "More and more, analytics services will deliver … easy to consume analytic applications in the cloud that will be very powerful," he says.

Traditional OLAP methodology will be enhanced too, and enterprises should expect more customisation and personalisation of information segmentations. To this end, the latest analytic capabilities will increasingly be geared toward high-volume forecasting - a greater possibility with more data streams.

Analytic processing is also likely to develop "a more granular edge", to avoid broad strokes with information that loses customisation possibilities for both end users and their customers. The third

area of business analytics evolution is the most opaque though. It is decision management support. In part "resurrecting the term decision support".

Automation and collaboration of the data behind this aspect of business is reserved for the leading edge of analytic maturity for the time being. However, as collaborative applications and semantic layers grow, the analyst says that there is promise for a boom in analytics to handle these typically manual yet critical business decisions.

An example of this would be the buy-in going on in the insurance industry space when it comes to claims processing, a manual but repeated process where brainstorming over systems is showing promise to provide personalised, and real-time responses.

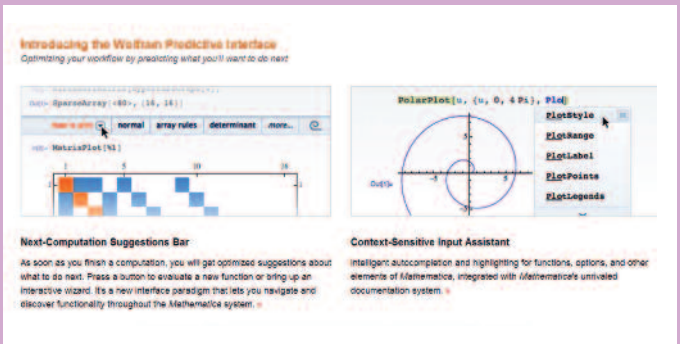
In preparation for these analytics changes, companies should be immediately prepared to map out pressing business problems and create a "finite list" of analytical enterprise teams. They should also create data discovery tool prototypes to address those business problems and subscribe an analytic industry service relevant to their business area. Within 12 months, companies should assemble teams to tackle high-volume forecasting and granular segmentation, and then, roll out support solutions.

AQ

Mathematica Upgrade

Nigel Cummings

Wolfram Alpha has upgraded its popular Mathematica application to Version 9, and surprise surprise, it now supports predictive analytics!



The latest version of the popular application has been equipped with 400 new capabilities, and a new emphasis on usability and automation. The automation and data science focused technology integrates computation into complete workflows.

A highlight of the new application is Wolfram's "Predictive Interface", a collection of intelligence works which as acts as combined software suite to "intelligently suggest" what to try next based on heuristics and data gathered from millions of Wolfram Alpha queries.

According to Conrad Wolfram, director of strategic development, "Even before 9, Mathematica was the broadest, deepest computation system in the world. [But] as scope increases, so do usability challenges. Our new Predictive Interface really helps. Getting Mathematica to think ahead means not only newcomers but all Mathematica users can access its power far more effectively."

Wolfram himself often talks of the need to "democratise computation" as his team works to optimise what's possible in each field of data science. Analysis and visualisation of statistical data and social networks (e.g. Facebook, Twitter) is fully integrated with existing capabilities such as instant interactivity, computable documents, and symbolic computation.

Mathematica 9 also offers formulas to ease social networking analysis, three-dimensional modelling and other computationally complex calculations. It includes a predictive analysis feature that provides users with suggestions of how their work can be further refined through use of the software.

Mathematica 9 is the first version of the software that allows researchers to analyse patterns of behaviour in social networks such as Facebook. A new function allows users to pull in data using APIs (Application Programming Interfaces) from social networks. The data then can be analysed and visualised.

According to Wolfram, the feature "promises to be very valuable not only for professional data scientists, but also for math and computer science students who want to jump immediately to the frontiers of one of the hottest current areas".

The Mathematica application provides access to a considerable repository of mathematical formulas, and Version 9 can also handle a number of new, tricky differential equations, such as equations with discontinuities; a model of a ball bouncing on a surface, for instance. More algorithms have been added for signal processing, control system modelling and vector analysis.

The purpose of these new algorithms is to reduce the amount of formulation researchers must do; computations in general relativity that even recently seemed like major research projects, now happen in mere seconds. This is also the first version of the software with built-in integration with the R programming language for statistics.

Mathematica also contains visualisation tools, and this release will allow users to view their data in three dimensions. It can export very large image processing tasks to more powerful machines, and provide live data visualisations that can be distributed to other users.

The new Wolfram Predictive Interface can suggest actions a user can take based on the context of the workspace. Another feature is the ability to calculate with units of measurement to its namesake software. Users can now add kilometres and miles together, or compare pounds to kilograms. The software supports units of measurement from a wide variety of sciences, including physics, chemistry, astronomy and engineering.

For the home edition, Mathematica 9 starts at £195 for a one-time license, or £95 per year. The standard edition starts at £2,035, or £815 per year and the enterprise version costs £5,695. Mathematica 9 runs on Windows 8, Windows 7, and Windows Vista; the Windows version of the application also works on PCs running Windows XP with Service Pack 3 or later. The application also works on any 64-bit Intel-powered Mac running OS X 10.6 or later. More information on Wolfram Mathematica 9 can be found at: **http://www.wolfram.com/mathematica/**

AQ

Artificial Intelligence Knows When People Tweet

Nigel Cummings

Salarix, the social media marketing company has upgraded its analytics product, Amplify, which is a machine learning program that focuses on social networks.

The purpose of "Amplify" is to search social media for conversations related to your business, but the application now goes beyond just looking for people talking about you. It looks for conversations that are relevant to the products and services you provide. Some of us may consider this rather intrusive analytics, but Salarix seem to think their 'improved' application gives users an even better opportunity to market products in 2013, both from you and to you!

According to Santanu Bhattacharya, Salarix CEO "Amplify" allows brands to build preapproved messages and target people that would be interested in them, and enables them to maintain control over messages and avoid "twitter fires". Apparently, this is the term given to a situation where one tweet sets off a sort of chain reaction a bit like the old chain letters, I guess.

For those of us interested in what Amplify has on offer, it's best described as an artificial intelligence capable of operating with "supervised machine learning." It bases its criteria for locating relevant conversations based on scores given to hundreds of thousands of tweets that have been read through by humans. Based on how the humans scored their relevance, the program can then use that as a basis for improving its performance. Customers can then continue to fine tune it as the program works for them.

Amplify is built around different types of product verticals, in recognition of the way people talk about things. "People talk different ways about different products," said Bhattacharya. "The way they talk about cars is different from the way they talk about insurance or electronics. Every industry is modelled to make it easier to figure out the tone of conversations and even identify sarcasm."

He says that context is crucial, too. “When people are talking about apples and oranges, the program figures out whether they’re talking about fruit or about tech companies”.

Of primary interest to marketers apparently, is Amplify’s ability to identify influential people talking about conversations where their product is relevant – and send them a targeted tweet. (You really know you’re important if a program like this singles you out it seems!)

Amplify even learns about what influential people are interested in hearing or talking about, so that someone who cares about the style of their car but not its fuel consumption for example, aren’t going to get tweets about how fuel efficient a car is.

By using artificial intelligence to monitor what people are interested in, Salarix believes that advertising can be made more effective – and less annoying. “Instead of just putting a banner in a network, Amplify sees what people are talking about before a message is sent.

AQ

Data Farming

Nigel Cummings

Not a new term, it has been around since the late nineties, “Data Farming” seems to be entering a domain near you now whether you like it or not! Why? Because it allows for the examination of whole landscapes of potential outcomes, not just a few cases!

Conventional data mining will see analysts expending a great deal of effort sifting through enormous amounts of existing information for useful nuggets of information, data farmers grow their own. Researchers typically ask a “what if” question, build a model and run a simulation thousands or even millions of times. Then they look for trends, anomalies and outliers in the “grown” data.

Data farming provides the capability for executing enough experiments so that outliers might be captured and examined for insights, yet data farming is not intended to predict an outcome; it is used to aid intuition and to gain insight.

It is a technique which helps analysts identify paths to success. It can for example outperform conventional modelling pathways; consider a relatively simple scenario in which the Red Team is protecting an objective and the Blue Team is trying to penetrate it. Perhaps conventional analysis allows for the discovery, after running the simulation a few thousand times, that Blue achieved its objective in only a few dozen runs.

But conventional modelling may only look at the cases where Blue succeeded in penetrating, and identified what factors enabled Blue to do so. Add data farming to the modelling and you could run simulations where Blue purposely takes that path rather than randomly, to optimise their chances for success. Researchers can also vary the inputs and then restudy the outputs, allowing better understanding of what affects certain things more than others.

Inevitably anomalous outcomes will crop up in such analysis, and these are often the interesting ones. Finding the cause of why a model behaves in an unexpected way may lead to the discovery of an error in the model, its assumptions or in one’s understanding of what is being modelled.

In defence modelling an application called MANA is proving useful to the New Zealand Defence Technology Agency, it can model asymmetric threats to frigates and arrive at a variety of interesting strategies. This may assist stakeholders in identifying the most useful tactical options, some of which would have required considerable lateral thinking to extrapolate from conventional analytical results. By adopting a data farming approach with multiple scenarios, figuring out where a tactic breaks down potentially allows analysts to gain greater understanding of tactical weaknesses and how adversaries might exploit them.

Data farming grows even more useful as problems grow more complex as the numbers of criteria and objectives increase. But a sufficiently fast computer can generate a vast array of possible outcomes and allow analysts to find the most desired ones when utilising the data farming approach.

Georgia Tech Research Institute is working with a data farming team to develop a socio-cultural model for a disaster relief scenario where shortages, lack of sanitation, famine and other factors could incite violence. The goal is to arrive at iterative processes to assist in highlighting emergent behaviour, where something that may seem like a good course of action eventually has unanticipated second- and third-order effects that actually make things much worse. Such insights would not be so easy to come by via conventional data mining and analysis.

Operational Researchers in Germany are developing data farming strategies too, because the use of the technique may prove useful for Germany's Federal Forces. Working in multidisciplinary groups is useful. In practical application, sometimes it is useful to divide large problems up for analysis. Klaus-Peter Schwierz, senior manager and O.R. advisor at EADS' Cassidian division recently said. "We are looking at the same question base under different angles of view, from a very high abstraction to highly detailed truth. We start evaluations in the high abstracted world, looking at the universe, and then we drill down to the highly detailed world. The two methodologies are not competing with each other; they are complementing each other."

However the adoption of new forms of analysis and data abstraction must not lead to paralysis through analysis, as with any modelling effort, data farming requires a balance. The model should be as simple as possible, and as detailed as necessary, with the goal being to apply the model, evaluate the results, and then visualise and present the results in a meaningful way.

Combined with global connectivity, data farming could herald a new era in reachback support. A commander in Afghanistan might for example send a "what if" to analysts back home, they in turn would number crunch using data farming techniques and send back insight on how to approach the situation as it unfolds.

The Swedish Defence Research Agency is already utilising data farming techniques to evaluate options for expeditionary operations. "We have found the data farming methodology allows for robust evaluation of existing equipment alternatives using existing or new tactics, techniques and procedures," said Johan Schubert, Deputy Research Director.

Data farming may also have value in improving acquisition decisions across coalitions by helping national leaders understand their options in aggregating coalition force structure to meet a range of threats. As you may have noticed, there is a trend in the application of data farming – defence! The military have always been "early adopters" of new technologies and analytical techniques, after all they generally have access to the most powerful computing platforms.

However data farming is likely to arrive at a desktop terminal near you very soon – where it may become the tool of choice in making modelling and simulation more effective for decision makers. Currently the most efficient data farming takes place thanks to the "number crunching capabilities" of super computers, but developments in office and even home PC technologies, point the way for ever faster, more powerful portable computing platforms accessible to all!

AQ