IBM Content Analytics Dialoginar

RANDALPH KAHN: Years ago, like at the beginning of the millennium, some IT guy gets tasked with building a database - a database, no big deal. He works for the Mining Safety Organization, but he's just a regular IT guy and he's tasked with building a database that's going to collect and house information regarding mining disasters. Makes sense. And so he does his thing and he creates the database and - let me stop there for a moment. Fast forward, the beginning of 2006, a couple miners are in a mine and an accident happens and they lose their life. And, as a result of that, the regulators and a whole bunch of other people look at the event and say, what happened? Let me go back to my database. You see, in 2003, the government accounting office looked at the Mining Safety's database and record keeping and said, hey, you're not getting it right. See, my IT guy wasn't properly tasked. They said go out and create a database, but, actually, this was a powerful thing. If I could know more information about these disasters, maybe I could avert the next one. Maybe I could understand what's happening and trends, but you, the IT guy that structured this thing, you didn't give me the chance. What happened? You see, the way that my IT guy set up the database, what he did was he allowed me to track accidents by mine, but he didn't allow me to track accidents based upon the owners of mines. But let me leave that for a moment, and I will come back to it. Fast forward, 2010, down in West Virginia, 29 miners lose their life in a terrible, terrible mining accident. And you ask yourself, is there anything that we could have done to know more about what ailed the company - oh, I forgot to tell you, the 2006 accident in which two miners were killed, and the 2010 accident, it was the same company. What could we have known to avert a disaster, anything? As it relates to the information that we have, whether we're agency or we're company, how do we best manage the information so that we can avert the next disaster? My mission today is to demystify and explore content analytics. Information is growing at such an incredible rate. Employees are ubiquitously mismanaging stuff. I've sort of come to this conclusion that to allow employees to continue to manage and find what we need to find for business problems or deal with properly locking down stuff for privacy or information security, or even responding to litigation, really is a challenge if you expect to have the employees do that. And combined with this vast, vast volume of stuff that truly, if it's an asset, you better harness it, you better look and find its value, I've come to conclude that we need to use technology to better manage the output of technology; namely, information. And I guess the question is, you know, to you, do you think that the content analytics technology is there, in terms of its sophistication, to solve those problems today? And then do you think companies are ready to harness that technology?

CRAIG RHINEHART: Well, I think the technology is there, and I think companies are ready, but it's really not the employee's fault, if you think about it, to go back to the first part of your question. I mean most

employees, it's not their job to manage information, it's their job to bring customers on board or close sales or process claims. As part of that, because of this information explosion, they now have this increased burden of managing all this stuff. And, you know, we're unfairly asking employees to do all these things, so you get…you get what you get, and the reason that organizations are ready is because this information is not only growing at such volumes today, it's gonna get worse, it's gonna grow more and faster in the future, and if you don't apply technology to this problem, you're never gonna be able to make sense out of it, you're never gonna be able to understand, out of all this information that's available to you, which subset is relevant to you and then manage that properly, whether it's for e-discovery or whatever.

RANDOLPH: So in the content analytics world, when I think of it, I think of there's an algorithmic piece that allows you to discern patterns and themes and extract value, right? And there's also a linguistic piece, right? Can you sort of help us understand how those work together and what they do?

CRAIG: Sure. So, first off, you know, we're talking about unstructured information here, and the first thing you have to do is bring structure to that unstructured information. So if you think about it, a 5 is a 5 in any language. It might be expressed differently on the screen, but a 5 always means 5. When you get to unstructured language and parts of speech, you've got sentences…languages rather that go left to right, other languages go right to left and, you know, words mean different things in all these different languages. So the first thing that you have to do to try to make sense out of this large amount of unstructured information is bring structure to it. You have to parse the sentence, you have to identify what language it is, you have to understand the parts of speech and, basically, bring the same kind of structure to the unstructured world. Once you do that, then you can use analysis tools to…or algorithms, to use your word, as a way to understand what it means. What's this text telling me? Is this someone who's about to sue me? Is this a piece of information that might create competitive advantage to my business? Is this a piece of information on the web that's telling me that someone is happy with my product? Are they so happy that I should try to sell them something else? So this is the value of this technology. But, first, you have to start with bringing structure to the unstructured before you can do all the really cool algorithms or analysis that really allow you to leverage that information for some kind of business advantage.

RANDOLPH: So when you think about the value of content analytics, to the typical business today, is it your belief that it has value in the BI or the business intelligence world, as well as perhaps discerning stuff that needs to be locked down because it's private information, as well as being able to help solve maybe more economically, more expeditiously, my e-discovery conundrum? Does it really do all of that?

CRAIG: You can apply it to all those areas, and it solves different problems in all those areas. It doesn't solve the same problems in every single area, it's different, it's situational. But take e-discovery as an example, you're a lawyer, you know, think back on your cases, you know, how many human beings does it take in the e-discovery process to read through, mark up documents, put them in the right organizational structure so that someone else can come along and look at them again to see how relevant they are to the case? Does that just take one or two people, Randy?

RANDOLPH: Well, actually, it takes 100 lawyers, 'cause we're gonna need to bill and bill and bill and bill until we've billed enough on the case, and then we will have solved our answer.

CRAIG: Right, but if you're the guy paying the bill, and you understand that this technology can cut out a large amount of that cost, and still get the same basic set of results, you can accelerate the time to making a decision about whether this is a case that should be litigated or a case that should be settled, that's gotta be huge value. I know it's certainly a lot less cost.

RANDOLPH: Are you telling me that lawyers are gonna be out of work?

CRAIG: No, of course not. But I'm telling you that there's an opportunity to reduce not only the risk associated with the e-discovery process, but certainly the cost associated with the e-discovery process.

RANDOLPH: So let's talk about the risk piece for a moment. So when I hear risk, what I actually think is you're giving me a better mousetrap. Are you telling me that if I go the way of content analytics and apply that to the e-discovery problem that I'm gonna actually have a better result; namely, I'm gonna be able to hold up my hand and say we got the response and stuff, we got more response and stuff here, we got a better response to our request if I use these tools?

CRAIG: That's a possible outcome, but I think it's more likely I'm gonna tell you something that you don't know. The value of content analytics, if you think about it, it's not like search. Search is a critical business tool, but search presumes that you know what you're looking for. Content analytics, on the other hand, is an exploration of a large corpus or a large set of documents or information, and it's about understanding all the hidden relationships that are in there that you don't even know what to search for. This is about finding something that you didn't know, something about my business, something about the case, if you want to talk e-discovery. The sooner you find that information and can take action on it, that's real value.

RANDOLPH: I used to be a believer that employees had to really sort of know the content in front of them to properly manage it. After all, if you create content, the likelihood is you're in the best position to

know what it is, know its long-term value, know how to harness or extract value. And what I've come to believe is, actually, they never use it, they never do it, and the companies that…that pay lip service to this idea that information is an asset, and they do nothing with it or blow wholesale content in huge volume away, without regard to its value to the enterprise or the business problem or the legal issue, from my perspective, I think it wholly misses the point, which is if it truly is an asset, you need to harness it. And I guess I have become a believer over time that the content analytics tools are truly the path forward. Is it your sense that customers are open to using these kinds of tools to solve a business problem or a legal problem?

CRAIG: Well, they're not only open to it, they're doing it. You know, this technology is used in a lot of places today. I'll give you a perfect example. We have a medical device manufacturer who uses this technology to understand product quality issues in their products. They're looking at information that's coming in from their call center. They're looking at inbound e-mails, complaints, success stories, these kinds of things. They're looking at publically available information on the web. They're…the FDA has a website that reports on medical device failures in the market. They're looking at all this massive amount of information to understand are our products working in the market? Do we have product quality issues that are being reported on in public forums and in our own call centers? And, if we do, let's remediate those problems sooner rather than later. You see that in a number of areas. I think my favorite story is a major metropolitan police department is using this technology today, Randy, and they're using it to solve crimes. What they…the problem they had was in each silo in precincts, they had police reports that were locked away on paper, they had police reports that were locked away in electronic systems, and these systems were not talking to each other across the district, so they had no ability to understand on a citywide basis the trends and patterns of crime. They put in a content analytic system and, basically, unlocked the value of all that information that was previously locked away on paper, locked away on other electronic systems, and then started correlating what was happening across multiple precincts in a metropolitan area. Within the first week of deploying the system, they solved two murder cases that had been cold for a number of years because they found something in the police reports that was not previously available to them, and they understood that there was a guy out there who had a tattoo on his neck that said a certain thing and that it was causing a certain crime. And they were able to locate that suspect, arrested him and he's in jail now. And there are many more examples of that kind of thing, that's just two.

RANDOLPH: So for the typical large organization today, is content analytics a nicety or a necessity?

CRAIG: It's a necessity. It is the only way, from a technology perspective, that you're going to be able to deal with these massive amounts of information and bring understanding or meaning to them.

RANDOLPH: One of the things that I've been thinking a lot about, as it relates to our clients, is how to cost justify new technology tools that allow you to better manage information and, as it relates to the content analytics, I see so many ways that you can cost justify an acquisition of a piece of technology. The return on investment is almost immediate, whether its in the context of discovery or a business intelligence problem. Are you finding that, in a down economy, that leading with a content analytics tool actually is something that folks are open to because its going to help them reduce costs and also be faster, better, cheaper and legally compliant?

CRAIG: Well, in a down economy, in any economy, you know, being able to help a customer reduce costs is a good thing, and that gets people's attention. Obviously, the problem is more acute in a down economy and it's an important part of that conversation. But I'll tell you a story to make it kind of real, a large financial services customer of ours basically came to us with two problems. This was in the middle of the banking crisis. The first problem was they were about to inherit the assets of a failed bank and they were gonna, basically, get a whole lot of information and systems that went along with it, and really had no idea of how they were gonna manage, you know, these systems and assets, this content, if you well. Yet, as part of that process, they were gonna have to, you know, preserve access to that information for a period of years, I believe its ten years, in the event, you know, record keeping purposes and so forth. So that was one part of their problem. The second part of their problem was employees who left the company. They had begun a practice of every time an employee left the company, they would make an image of the laptop or the computer that the employee was leaving, don't destroy the files, let's just pile them up on a file share somewhere. So this keeps going on and going on and going on. Meanwhile, what you end up with, in the case of this bank, was over 100 petabytes of data, no one knew what it was, piling up, eating up power, paying software license fees for applications and servers and systems just to keep stuff going because they had an obligation to keep it around. Why on earth wouldn't you take a technology like content analytics, that can go through those piles of information, tell you what you need to keep, because you have some obligation to keep it, and dispose of, decommission…defensively decommission all the junk that's in there that you're paying all these fees for, you're paying to keep the power on, you're paying these software license agreements? There's no reason to keep hanging that around. The reason people keep it hanging around, at least in the case of this institution, was they couldn't separate the unnecessary information from the necessary information, and that's what this can do. You can immediately go through, you know, these piles of information that have crept up over a period of years and separate necessary from unnecessary. And when you think that…when you put some math to that, and the industry estimate is 70% of the information that's being kept today in enterprises is already past its retention date, so why are we still keeping it? Big numbers, big numbers.

RANDOLPH: How does content analytics differ from auto-classification technology?

CRAIG: Oh, great question. They're similar, but they're not exactly the same. Auto-classification technology does exactly what its name implies, it automatically organizes or classifies information, content, documents, whatever it might happen to be, into some sort of structure. So if you've got a filing system or a taxonomy or a file plan, whatever…whatever you call it, and you want to put documents into that structure, or any structure for that matter, this is a way to automatically do that, without having to have, you know, the human beings in the business process make these decisions. It's a technologically approached…a technological approach to classifying or organizing your stuff.

RANDOLPH: If I'm a business person and I don't have any technical background, but I've got a whole host of business problems and I want to use content analytics to help me solve them, is there any way that I can take that in-house and, me, a nontechnical person, make this thing work for me?

CRAIG: Content analytics, you know, can come in a variety of flavors. Our version of it, what we deliver to our customers, has a very strong out-of-the-box simple user interface, so, immediately, on day one, you can start getting value out of analyzing and understanding your…your content. What's not gonna happen is a bunch of guys with lab coats and protractors and slide rules aren't gonna show up wanting to sit with you for weeks and months and perfect your BI environment or your analytics environment before you ever see day one of value. You'll start getting value out of this from the very, very beginning.

RANDOLPH: There's no question that content analytics can be useful to cull through and organize and understand vast quantities of stuff on the web, and I definitely get that, and to the extent that you're able to glom onto information say, for example, customer complaints, or really understanding your customer, it would seem to have incredible business value. Can you give me an example of a client who uses content analytics and extracts really the user experience so that they can make a better mousetrap, a happier customer with that mousetrap, do you have such an example?

CRAIG: I do, I do, but, before I go there, let me point out that the web is a big problem when it comes to this, and by problem I mean its size is a problem. There are so many tweaks, there are so many postings, there is so much social information being generated that, unlike inside the firewall, you need a significantly scalable approach to solve the kinds of problems of the web, just by the sheer mass of information. Obviously, the larger the dataset you're working with, the harder and the longer it takes to run these kind of analysis. So, obviously, this is a problem that we think that IBM is particularly well-suited for, you know, not only do we understand these kinds of problems, but, you know, we have a reputation for being able to tackle problems of this kind of magnitude. Now with that being said, that's precisely the reason that NTT Docomo came to us a number of years ago, because what they wanted to do was reduce customer turn by having a deeper understanding and a better understanding of what their customers wanted before it was too late, after they left. So, in their case, we've been working with them

for a number of years on a voice of the customer project, and they do everything from understanding web content, tweets and Facebook and this sort of thing, along with the notes and input from their call center applications, inbound customer e-mails, as well as recorded conversations that have…that are then converted to text through speech-to-text conversion technology. They analyze all of this text and are able to do things like introduce new offerings before they start losing customers, new pricing plans, new phones with new capabilities, you know, new services in new areas. And they've introduced a number of offerings that have…that have enabled them to take market share from their competitors and put a significant reduction into, you know, loss of customer, customer turn. So it really doesn't get any better than that, when you think about it, 'cause what they've done is they've reinvented their relationship with their customers into a more of a I understand your problem as you're starting to experience and I can now take action before you're, you know, you would ever even think of leaving as a customer. So it's really one of the best examples we have of how this technology can help, you know, generate new streams of revenue for your organization, which is exactly what NTT Docomo is using it for.

RANDOLPH: Most folks think about content analytics in the business intelligence or BI space, and I understand that it's quite useful for a couple other things. Let's say, for example, we have litigation, right, and we want to really both minimize the exposure, minimize the inconvenience and, therefore, the expense to the organization by wasting employee time, we want to give the best possible response, and also maybe understand our case a little bit better. From my perspective, content analytics can be very useful in that regard. Do you have any thoughts you'd like to share about that?

CRAIG: Well, I agree with you, it can be very useful in that regard. Everything from prior to collecting the information that's relevant to the case, culling it down and understanding it so you can start to make decisions about the case, and then, ultimately, sending the smallest possible dataset off to your hosted review provider. Content analytics helps in all of those ways because it allows you to make decisions earlier in the process, see what's out there, assess are you in a good position, are you in a bad position, is this something that you want to litigate, might want to litigate, you know, don't want to litigate, it really, really accelerates that time to decision from a legal perspective, as well as allows you to work with a much smaller and more relevant set of information when you start going through the review processes of the documents and the information to the case. Everybody over-collects. Everybody goes around to all these information sources and, out of this sort of fear-based, I don't know what I have, there might be something in there, I don't want to miss it, everybody over-collects. Well, what does do? It means you spend more time and more dollars reviewing documents that really aren't necessary to review. So let's look at that earlier in the process, let's use content analytics and only collect what's relevant to the case, and let's not make decisions from a position of fear, let's make decisions from a position of insight that we know what's out there and let's collect what's relevant, and we benefit because we take cost out of the process and we take risk out of the process and we're able to make better decisions and faster.

RANDOLPH: Records management is a topic that gets a lot more play these days, but one of the things that's clear to me is that employees are not really keen on doing it, and if they're doing it, they're not doing it particularly well. They've got a full-time job anyway, and with the economy as it is, they're…the people that run their business likely, you know, does not want them to be spending half of their day coding e-mail. So my opinion is that on a going forward basis, whether it's the classification technologies or it's the content analytics technology, it's, again, remarkably useful at dealing with the records management and the classification problem that every single business experiences. Do you agree with that? Do you disagree with that? Where do you stand?

CRAIG: Well, it's useful in two ways. I agree with it, it's useful in two ways. The first is on the front end, identifying what records you have, where are they, are they where they're supposed to be? I mean we're not talking physical records here where, you know, once every six months we have records cleanout day and everybody, you know, packs up the files and they get barcoded and palletized and go off to our friendly storage facility where you know where they are, they're safe, you can go visit them if you want, you can get them back, you know, we're talking about information that, you know, propagates on its own. I had a customer call it information rabbits one time, right, you can forward an e-mail to 100 people in the blink of an eye, and this creates a whole new class of problem and creates…this contributes to this explosion of information, you know, the propagation of stuff, duplicate. You know, I think I saw a stat just the other day that said an e-mail with attachments are stored, on average, 12 different times. That's a records problem. How on earth are you gonna find all these duplicates? Technologies like content analytics allow you to go out and look at vast corpuses of information, figure out what's out there, are they records, yes/no? Are they in the record keeping system, yes/no? Let's move the stuff from where we don't want it to be to where we do want it to be and have the proper controls and retention and so forth put on top of it. The second piece of that is this an audit tool. What a brilliant audit tool. I know I've read in your books the importance of having auditing as part of a record keeping program and what can happen if you don't audit. Well, you can use this technology in the same way. We have a policy, we said we were gonna do these things, we said we were gonna use a record keeping system, let's go back and check and see if we did. Let's audit ourselves. Let's go back and look at those same environments. Are our records still in the right place? Has rogue or bad behavior propagated up and we've got some records that aren't in the record keeping system? This is gonna find them. This is gonna allow you to remediate those kinds of issues before you have to deal with it in a much more punitive scenario, like when a subpoena shows up.

RANDOLPH: Hey, Craig, it was great to talk to you today. Thank you very, very much.

CRAIG: Randy,, great to talk to you as well. It's always fun when we get together.