Web Mining will Make Google Look so Jejune

DAVID CARPE, *Clew, LLC*

Leading research minds in Boston conclude that "the internet is getting wicked huge."

Taming and training the information beast are critical next steps toward delivering a more useful global information repository. While standards-making bodies like the World Wide Web Consortium (W3C) promote further advancement of agenda items like the resource description framework (RDF, a lightweight ontology system to support the exchange of knowledge on the web), others look far beyond these standards toward a new era of web mining.

I had an opportunity to talk with Mr. Laurie Lock Lee. Laurie is a principal knowledge management consultant at CSC in Australia who has just published a major grant-funded research paper about web mining (WM).

After reading this work and speaking with Laurie, it is easy to imagine a future where professionals discuss electronic research like popular film (as in, "my search results were so two-dimensional, and the results were poorly developed – not believable at all"). What follows is wrap-up of this conversation.

Can you tell me how you got into WM and what you do now at CSC to pay your bills?

I originally trained as a metallurgist with BHP Billiton – the world's largest diversified mining company, headquartered in Australia. I spent 10 years in their corporate research laboratories leading their IT research area. In the 1980s we focused on artificial intelligence (AI), knowledge based expert systems, and advanced human computer interfaces.

My interest in knowledge based systems evolved into knowledge management (KM) in the mid 1990s. Since my metallurgy days, I studied computer science and had an interest in how research gets commercialized. After a big gap, I've enrolled in a doctorate program in business administration to pursue my research interests in the relevance of social capital to intellectual capital and intangible asset management performance.

You've been involved in the past with SCIP, are you still?

Yes, I have presented at SCIP in Australia. I had specialized in the development of communities of practice, which are very similar to *shadow teams* in SCIP parlance. I got to know Babette Bensoussan from Mindshifts and met Ben Gilad, author of *Business Blindspots*, when we invited him to meet with BHP. I haven't caught up with the SCIP folks recently.

Before discussing your report on web mining, could you perhaps offer a simple definition of WM?

One could define web mining as simply discovering interesting patterns of information on the web that might provide you with some useful new insight, like a new market or a new consumer trend.

Web mining is really an extension of data mining. The source is largely textual, whereas the source for data mining is usually numerical. Data mining has its roots in machine learning, which is a sub-area of AI research. The aim of machine learning is to use computer systems to learn new patterns from raw data.

The competitive intelligence and KM communities are providing the context for where web mining can add business value. The standards groups are providing the enabling mechanisms . . . we need both.

A lot of attention is being paid to the semantic web, including such areas of evolution as the RDF from W3C. How will more structured data change forever the way WM works?

The semantic web should make web mining even more effective. In essence, the aim of the semantic web is to make the raw information found on the web more *machine readable* by providing metadata to help with the interpretation process.

The semantic web's dream is to apply this codification to all internet information. What is the potential impact when some information is accessible via WM built on standards such as RDF, and the rest falls far behind?

Of course that is the semantic web dream. At the moment we are only tapping a fraction of the information available on the web. I don't think it will come quickly though, and it's unlikely that we will be able to backward engineer all of the current information on the web to meet the semantic web standards. So it will be an evolution, rather than a revolution.

Major information repositories, such as government and universities, seem to be falling behind with standardization, and use designs and layouts that can't even be seen by all Internet

users across the globe, let alone searched. Does this concern you?

What this provides is an innovation opportunity. I can recall many instances of the industry decrying the slow adoption of standards like SGML, Core Graphics, MAP and others. Then along comes a de-facto standard that successfully balances functionality with pragmatics.

In-Q-Tel Ventures (a funding vehicle for the CIA) has made investments in the areas of KM, search and distributed data collection. Do you feel that government is driving change in WM?

It tends to put a sharper edge on the investment's need to succeed commercially, rather than just providing grants. In addition, the government is already using advanced WM technology for open source intelligence initiatives. When I worked in AI, much of the funding for AI research came from government sources, so you would expect that governments will have an early interest in this technology.

Your research also discusses WM applications such as bio-informatics. How WM might impact the field of competitive intelligence?

It is going to be increasingly difficult for companies to keep their intellectual property hidden from the forensic capabilities of new web mining techniques. Some of the examples I use in the report show how WM can help discover *human contacts*. As the SCIP folks will tell you, much of the useful intelligence is held by human contacts, so discovering who knows what (through social network analysis) is as important as interpreting the raw information.

Social Network Analysis (SNA) differs from what many call *online social networking*. SNA as a discipline is about analyzing existing socially derived networks, looking to identify patterns of relationships that might lead to improvement opportunities.

On-line social networking software applications like LinkedIn and Friendster are not really looking to analyze existing networks; they are trying to use them to create an on-line community space for networking. In some ways these community spaces could in the future become a source for an SNA, perhaps exploring the effectiveness of on-line networking.

You break the world of WM products into four distinct universes: business intelligence, customer relationship management (CRM), analytics and search/ meta-search. Which is the closest to addressing the WM needs of tomorrow?

The core technologies will come from the analytics and search areas. CRM and BI will be adopters of these new developments.

If you could command sites like Google to make one or two big changes, what might you require?

Clearly textual clustering or summarization, which is what we are seeing from companies like Vivisimo, Autonomy, Semio, and Verity. This technology on top of a Google-type search engine, which reaches even further into the rich data sources available over the internet, will make the web even more useful than it is today.

There are other research centers on the cutting edge of WM, particularly Carnegie Mellon and the IBM Research Labs. Carnegie Mellon has a strong heritage in AI and natural language research. Likewise with IBM; they have recently embarked on a joint venture with Factiva to develop a huge machine-readable, XML enabled text repository called Web Fountain that IBM is hosting at its California laboratories.

Is scoring or rating information a part of any vendor or research group's prerogative?

The ability to effectively *sift garbage* is what will separate good WM technologies from poor ones. I would expect that a good WM technology would weight information source credibility within its algorithms.

Are there major trends in WM standardization that you think people should be watching more closely?

XML is already the de-facto standard for improving the ability for machines to interpret textual repositories. I would anticipate that XML will become as pervasive as HTML. We in fact may still see a de facto standard emerge. Recall that HTML blasted onto the landscape when the standard setters were still playing with the HTML superset SGML as a hypertext standard. And who cares about SGML now?

Your research brings up the concept of WM oriented machines pre-processing information for people, incorporating what sounds like dynamic personalization. Could you expound on this?

I think what you are referring to is the smart alerts. The idea here is that many information services provide alerts on particular events to their subscribers. It is left to the subscriber to integrate the alert into the pattern of previous alerts or events to assess whether an action should be taken.

The idea of the smart alert is for the system to keep track of these temporal events which would enable the system to provide a richer alert or emerging trend to the intelligence consumer. I guess it's the text equivalent of a numerical trend.

Should information always be free and universally accessible? If powerful WM tools create a very transparent society, will the potential for abuse become a real consideration?

I think there will always be proprietary tools that will provide a competitive advantage over the *free to air* offerings. This is not unlike

comparing subscription information services with free information sources.

Clearly we can expect government privacy regulations to rein in inappropriate use of WM technologies. This is already happening. . .probably more observations than examples. Particularly with public institutions, the issue of privacy invariably is brought up when discussing potential web mining applications, in particular those that mine email repositories.

Mr. Laurie Lock Lee is a principal knowledge management consultant with Computer Sciences Corporation in Australia. For the entire report, contact him directly at llocklee@csc.com.

David Carpe is the principal and founder of Clew, LLC, a competitive intelligence consulting firm serving several of the world's most formidable organizations. He is also the founder of PassingNotes.com, a research community. Before selling out to pursue a career in business, raise venture to start a software company, earn an MBA, and create Clew, he earned a BFA in studio art. David resides in Boston with his two sons and their one-eyed dog. He may be reached at david@clew.us.

MCGONAGLE— *continued from page 51*

Competitive Intelligence, Quorum Books, p. 77-89.

Gordon, Ian (1989) 'Analyzing competitive intelligence,' in *Beat the Competition!*, Basil Blackwell, p. 111-155.

Fuld, Leonard M. (1995). 'A practical approach to analysis: analytical techniques and cases,' in *The New Competitor Intelligence*, John Wiley & Sons, p. 359-414.

Hussey, David and Per Jenster (1999). *Competitor Intelligence: Turning Analysis into Success*, John Wiley & Sons.

Kahaner, Larry (1996), 'Analysis,' in *Competitive Intelligence*, Simon & Schuster, p. 95-132.

McDowell, Don (1998). 'Selecting analytical approaches,' in *Strategic Intelligence: A Handbook for Practitioners, Managers and Users*, Istana Enterprises, 176-193.

McGonagle, John J. and Carolyn M. Vella (2003), 'Tips on managing analysis,' in *The Manager's Guide to Competitive Intelligence*, Praeger, p. 135-144.

McGonagle, John J. and Vella, Carolyn M. (2000). 'Analysis,' in *The Internet Age of Competitive Intelligence*, Greenwood Group, p. 97-126.

Tyson, Kirk W. M. (2002). 'Analyzing your competition,' in *The Complete Guide to Competitive Intelligence*, Leading Edge Pubs., p. 10-1-10-17.

John J. McGonagle is managing partner, The Helicon Group, Blandon, PA, a competitive intelligence consulting, training and research firm. jjm@helicongroup.com