



Mining the Business Intelligence from Unstructured Information

By William Laurent

Growing enterprises continue to watch their internal corporate data and information proliferate at unmanageable rates. Unfortunately, much of this information remains unstructured and difficult to catalog and index, let alone attach or associate meaningful metadata or contextual classifications so that the data can effectively drive business intelligence (BI). The overall picture becomes more clouded when organizations start to leverage outside data, much of it from the World Wide Web, to augment their BI processes and data sets. The bottom line is that there is a veritable wealth of intelligence available on the public Internet that can provide distinct competitive advantages to companies once they learn how to harness it.

Businesses are asking two critical questions:

- How do I get meaning out of all the unstructured text that presently exists in my company?
- How can I capture, process and use information that is freely available on the World Wide Web to be more competitive or enhance business operations?


These two dynamics are on a collision course. For unstructured information woes, life is getting better. Several applications are able to elegantly classify and categorize such data: document management systems have gotten much smarter and more flexible in the last few years, and XML standardization projects have been the catalyst for companies to better handle the integration between orderly and unstructured data. However, valuable data still lies dormant on various platforms where domain knowledge is exchanged, such as corporate email, company discussion boards or wikis. Furthermore, the irony for some organizations is that the most valuable data may lay outside their corporate boundaries - on the public Internet. This is especially true when it comes to threat management and risk mitigation. Looking at threats from an outwardly focused perspective has been a necessity long before Hadrian built his wall across the width of modern-day England. In keeping with this theme, the Internet remains the world's greatest untapped source of BI.

Law enforcement professionals have recognized for some time the wealth of valuable information that exists on the Internet. Likewise, chief security officers (CSOs) are increasingly using Web-based (as opposed to internal) data to help formulate business strategy and aid them in buffering corporate assets from a broad base of direct and nondirect threats. In both the government and commercial sectors, threats to business continuity and efficacy are always changing and mutating; enterprise data will always lag behind real-time Internet-based data, which is being created, updated and deleted every second of the day. Many risk managers now operate on

the assumption that what you don't know is more important than what you already do know when it comes to confronting risk. For this reason, many in the CSO and chief risk officer (CRO) community consider the Internet the "source of first resort."

Because CSOs and CROs often start their careers in the public sector, they tend to intuitively recognize the Internet's implicit value-add in matters of business continuity and due diligence: "What activist groups are out there wishing to harm my organization? Is there some recent violation of business ethics associated with a potential partner or investment candidate?" Sometimes this information is readily available with a simple search, but it takes time to store and categorize things in a meaningful manner. Therefore, a new genre of tools that can dig deeper into the Internet has emerged. The greatest benefit of this type of software is robust archiving functionality. Information on the Internet is much more transient than many of us realize. Many times, Web-based data is here today and offline tomorrow, so having the capability to properly save data sets becomes a critical requirement.

Kirk Dauksavage, CEO of RiverGlass Inc. (a provider of intelligent Web monitoring software) observes: "Once we get an engagement going with the CSO or CRO of a company, it is not long before they comment that we should be talking to some other department in their organization because they too could benefit from a solution that effectively locates and summarizes relevant business information off the World Wide Web." Being able to sift through mountains of unstructured text, mostly critical external Web content, is quickly becoming a necessary means to robust risk management and competitive intelligence. It will be difficult to have a sound information management and overall BI strategy for the future without addressing this imperative.

The need to glean intelligence from unstructured and loosely managed data will be driven by all business segments - marketing, sales, HR, finance, operations, etc. - as they look to increasingly make business decisions from knowledge contained in this data. Data which was previously thought to be unclassifiable or unmanageable is getting renewed focus now that its value is understood, and software exists to help make sense of it. Technology directors will be increasingly tasked with integrating and merging various types of mutating external data with internal relationally structured information. Such projects will give business leaders fresh perspectives on their existing data, creating a paradigm shift of how BI should be supported and competitive advantage can be sustained. 

William Laurent is a renowned independent consultant in data, governance and IT strategy. Please contact him at wlaurent@williamlaurent.com.