Data and Text Mining:

Event Horizon or Sunset?

A Talk Prepared
by Stephen E. Arnold
for Infonortics ICIC 2007

October 22, 2007

Data and Text Mining: Event Horizon or Sunset?

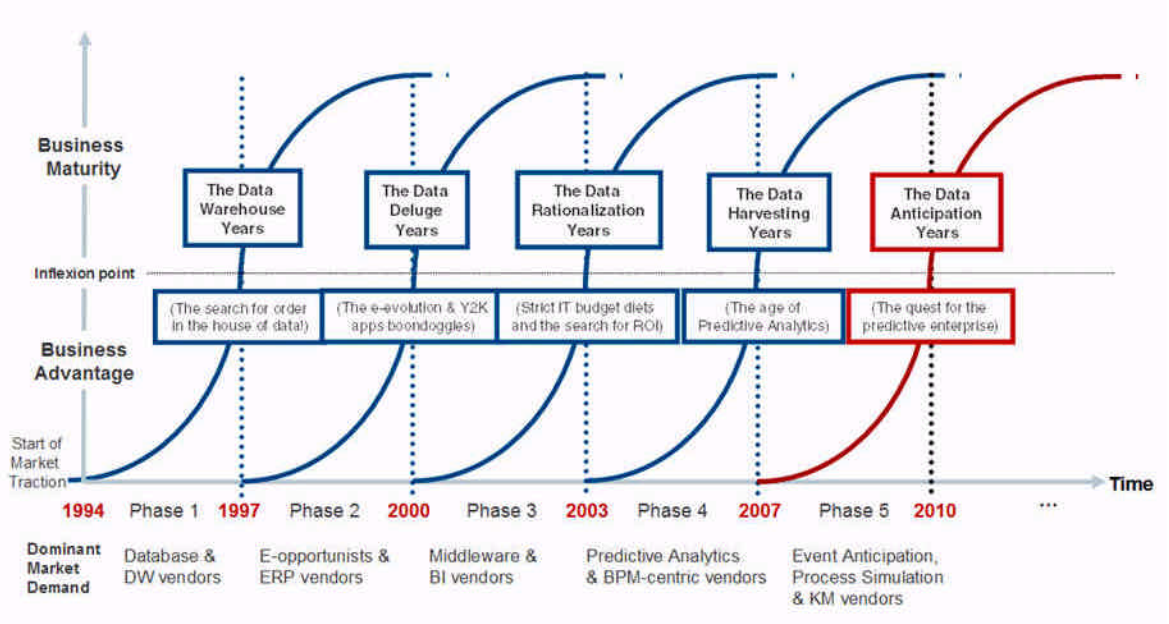
1.0 Hunger Now

Data and text mining have been of enormous value in some business sectors. In order to make data mining (extracting "facts" and "insights" from structured data) and text mining (extracting "facts" and "insights" from unstructured data), significant resources are required. In industries like pharmaceuticals, these massive efforts which can cost tens of millions of dollars consist of proprietary information and much larger volumes of open source data.

Executives have been willing to invest in content mining because it makes intuitive sense to "extract" as much value as possible from "information resources". The jargon used to discuss content processes borrows heavily from hard-rock mining. Indeed, many content mining professionals talk about "nuggets" of information.

Today, however, nuggets are not enough. We have entered into what SAS, the large business intelligence firm in the United States, has called "the data anticipation years". The goal of executives is to take "nuggets" and convert them into even higher value content objects. This is the business equivalent of saying, "Okay, you've mined the ore, and you've produced gold ingots. Now we want gold tiaras, watches, and necklaces."

The idea is that the processes of content mining must do more, much more to justify the huge costs to drug companies, military intelligence agencies, and financial services firms who live or die (quite literally) from information.



2.0 State of the Art

Much emphasis is placed upon the complexity of certain types of information. A casual Web surfer who encounters content in the form patent documents, gene sequences, or chemical formulae would say, “No problem. I read this type of material every day and with gusto.”

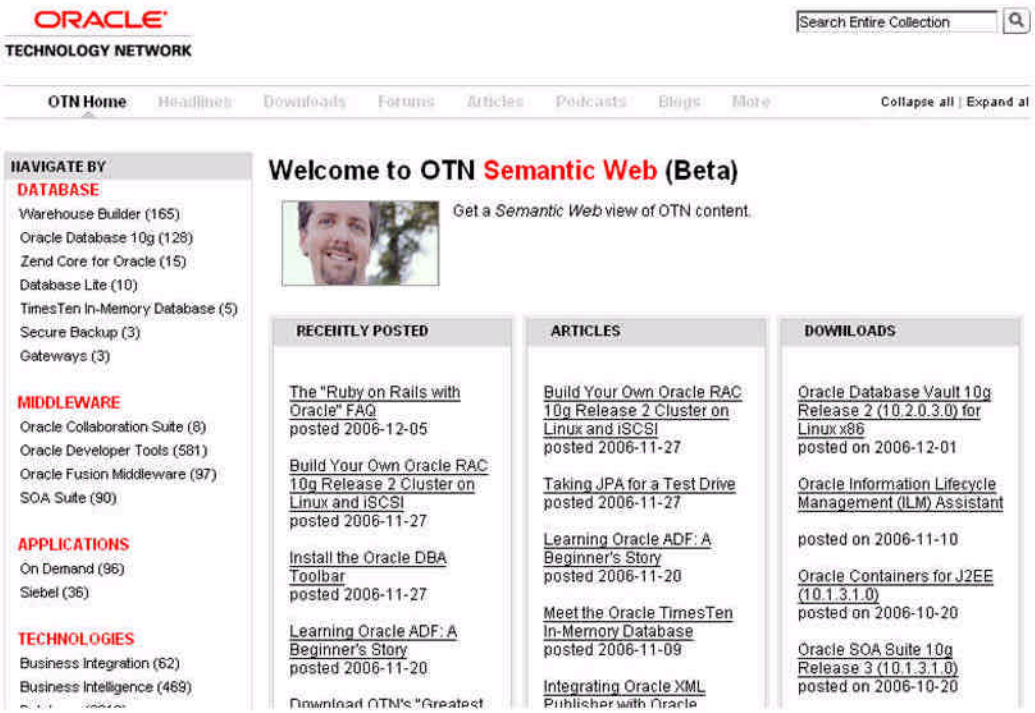
But scientific, technical, and medical information are also among the easiest types of information to process and manipulate. First, the jargon used in these documents constitutes a distinct language. A Ph.D. is certification that its holder can understand a particular field’s argot. Medical terminology has been beaten into submission over the decades by the commercial publishers, standards-obsessed governmental agencies, and commercial enterprises who tame very narrow, very specialized bodies of information.

Second, the equations, formulae, and diagrams are now more easily managed. There are the intricacies of math XML, ASCII representations of chemical structures, and strangely hypnotic sequences of letters to represent genetic information. With time and effort, it is possible to take a document chock-a-block with Latinisms, equations, and drawings to index each component and make most of the document’s components available for digital manipulation. Innovations from Adobe, InfoPrint, and Microsoft, among others, are going to make well-structured, compound documents more accessible each day forward.

But the problem is not technology per se. The problem is that modern research requires more than ever deeper, more recursive analyses of “what we know”. The issue is “what don’t we know”, to paraphrase one of the George W. Bush administration’s most colorful war fighters.

Companies like Oracle tout their content processing technology. But when it comes to making the data locked in large Oracle databases using Codd technology that is not decidedly long in the tooth, Oracle’s own tools don’t do the job. For example, here’s the interface to the technical information available on the Oracle Technology Network. Users are engineers, certified developers,

and Oracle employees. Notice that this interface is very different from the command line and the PL/SQL SELECT command that Oracle thinks its customers know and love:



Notice that this interface processes content, discovers categories, places the content into an index, and exposes that content in an intuitive, obvious interface. The user doesn't know what he needs to know before running a query. So, this interface makes it easy for the user to learn what's available, what's new, and what types of information are available to him. Oracle's vendor of choice is Siderean, one of the "beyond search" vendors who are responding to the new demands for content processing.

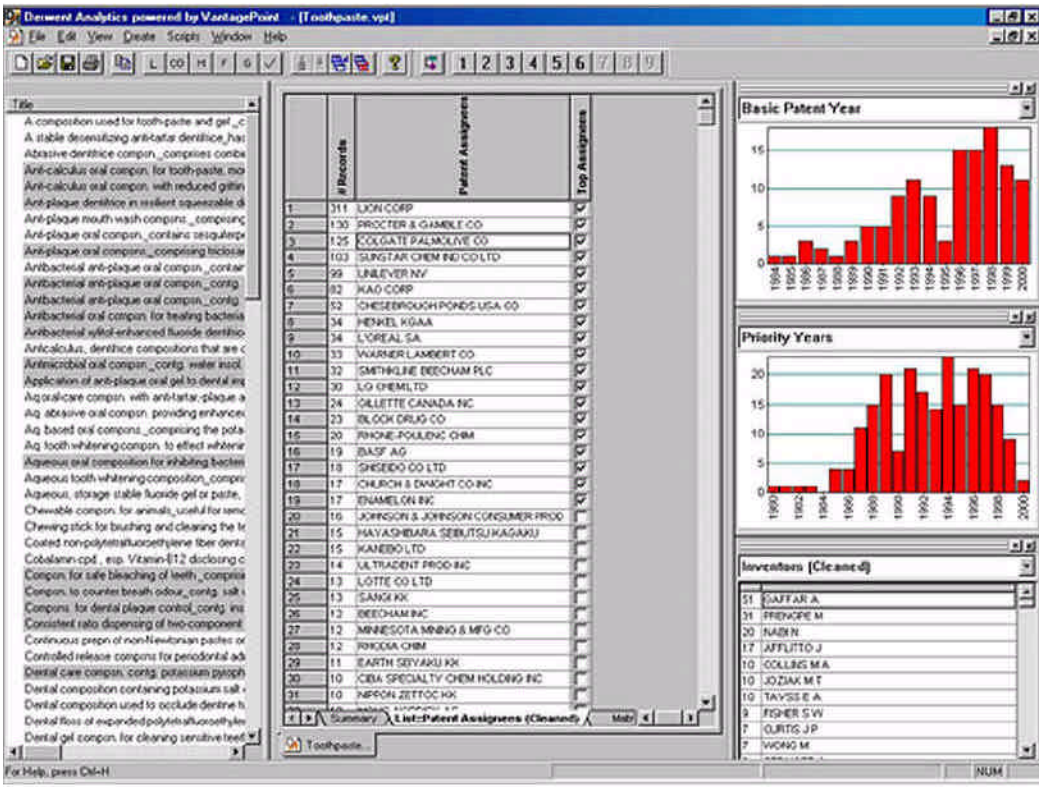
The content may be less "complex" than a gene sequence, but I'm not sure that programmers would agree. The Oracle programming and architecture data are a blend of algorithms, programming rules, third-party accessible application programming interfaces, and data.

The key point is that Oracle's own tools don't do the intelligence job needed by Oracle's own users. The situation is far from unusual. Most of the vendors of content mining tools present one picture in their demonstrations and marketing collateral and quite another within the software environment itself. "Sell the sizzle, not the steak" is the motto of some content mining vendors.

3.0 Specialist Interfaces

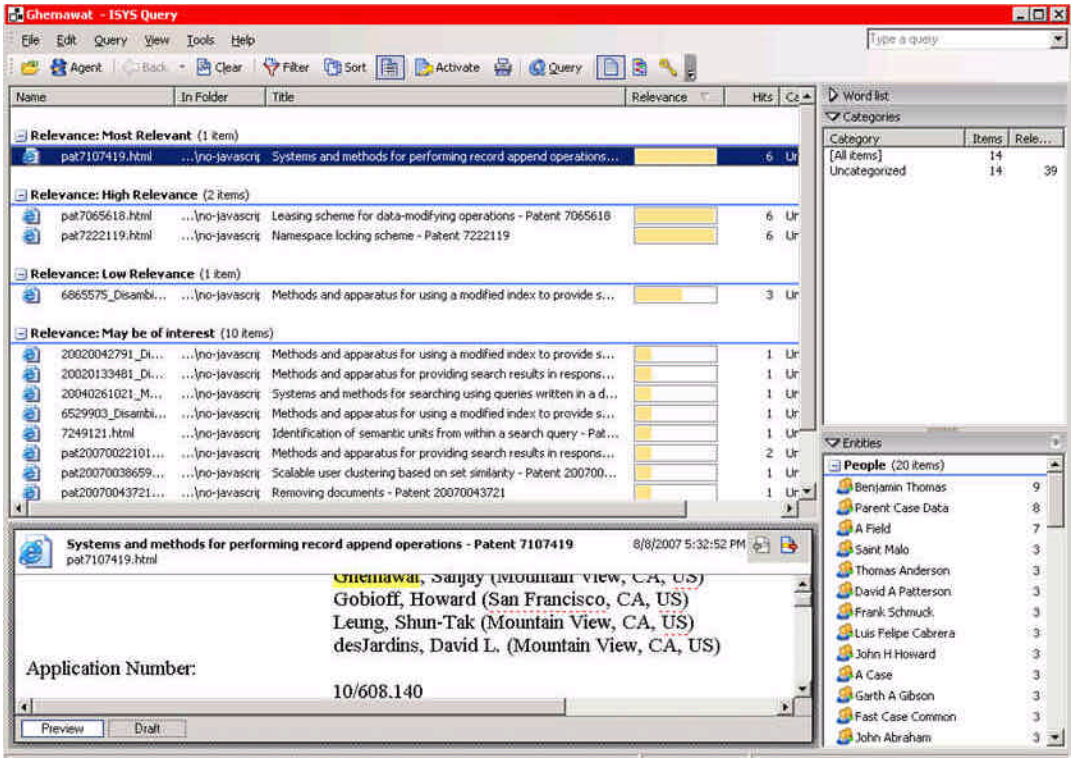
Companies involved in commercial databases offer powerful tools to process, sift, and analyze content. Here's an example of an interface from the giant Thomson Corp. My understanding is that this interface is highly valued by many researchers in scientific, technical, and medical

research. I am not certain, but I believe the U.S. firm Vantage Point provides some of the plumbing for the Thomson services such as Derwent that offer analytics.



The Thomson interface looks imposing. When I first saw it, I thought, “I wonder how long it will take me to learn how to use this powerful tool?” I also realized that its very complexity would ensure me of a job. No end user without special training and a keen devotion to a particular subject domain could make much sense of this display.

Now contrast the Thomson interface with the Siderean / Oracle interface or this interface from ISYS USA:

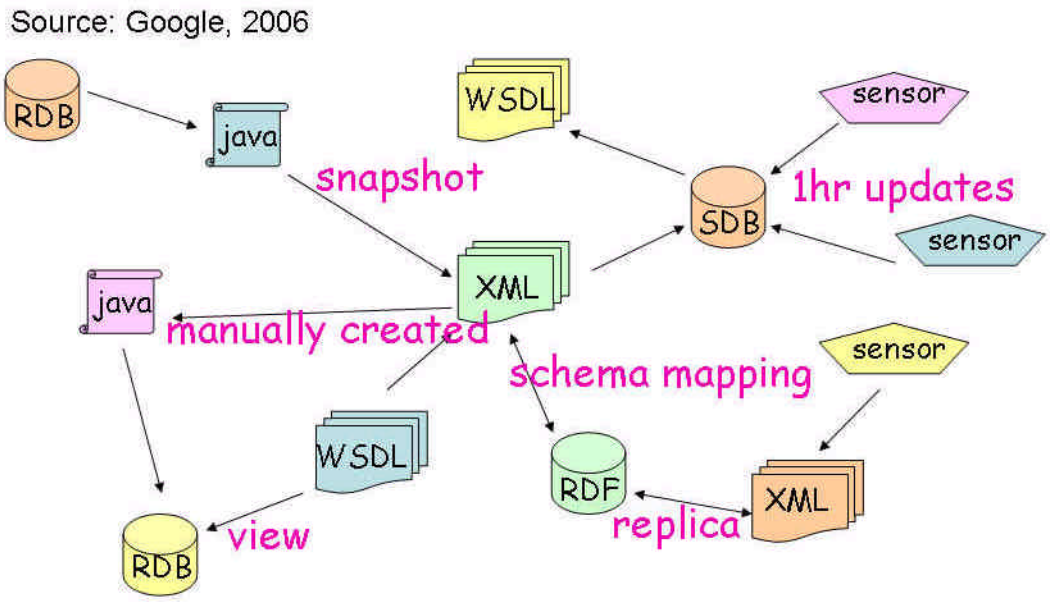


The interface is less cluttered and arguably less powerful than the Thomson interface. But the ISYS does provide clear indications of which document is relevant and it offers point-and-click access to content with key words highlighted. And, ISYS also includes a list of named entities. For a manager rushing to locate information, it depends on which interface is "better". The person trained in Thomson's approach will endorse Thomson and point to its options, multi-faceted analytic tools, and its content-charged interface. The arguably less technical manager may elect to use the point-and-click Siderean interface or the cleaner ISYS USA interface. It's like "love". Definitions are difficult to cement to a single Platonic notion.

4.0 A Broader View of Information

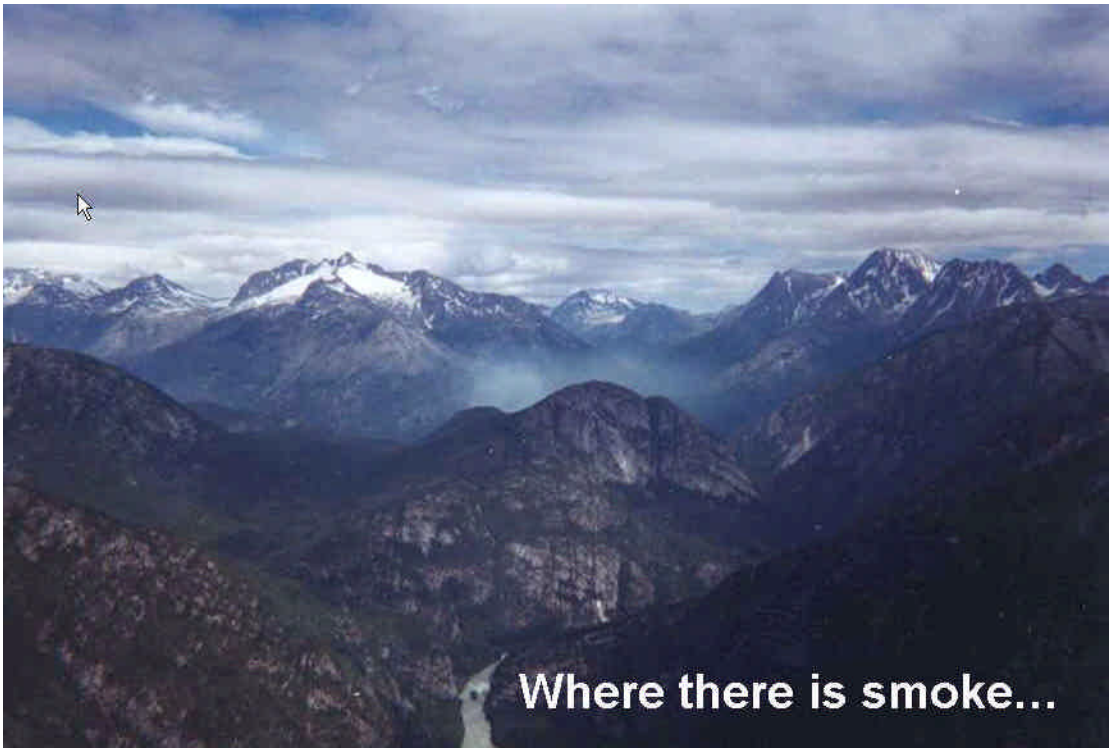
Searching on Google for information about content processing, I came across a series of slides prepared by Dr. Alon Halevy, now at Google. One of these slides appears below, and I am deeply grateful for Google for including this information in its cache. Dr. Alon Halevy's work can be difficult to find. There's not much available, and he seems to have changed his name from *Levy* to

Halevy in the last few years. I find this quite interesting because his work is seminal but tracking his articles down is a challenge.



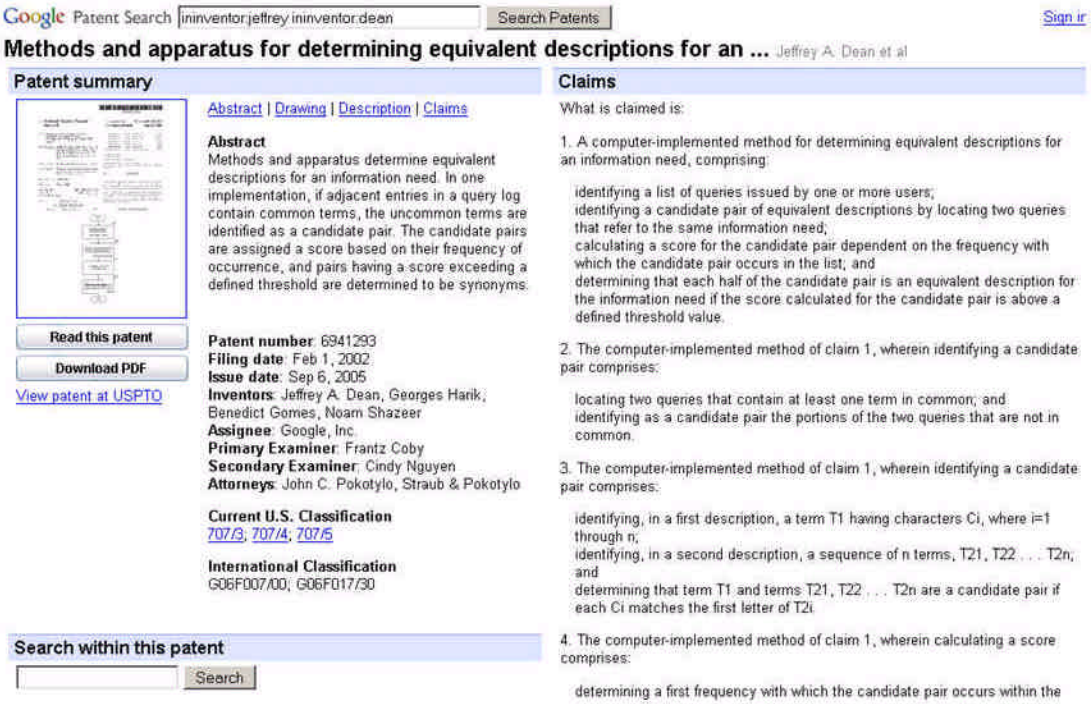
Based on my understanding of Dr. Halevy's work, this diagram shows the process needed to convert a number of different types of information from many different sources into some type of normalized representation. The implication is that content mining has reached the point that its usefulness increases when mining systems have access to a wide range of information. In short, I interpret this diagram as signaling that the era of knowing more and more about a very narrow corpus such as a patent collection won't deliver the "predictive" results that management expects. My query to Google on this matter when unanswered. But because of my new study *Google Version 2.0*, I am a PNG (*persona non grata)* at Google I have learned. Google it seems has an expert looking at making large volumes of content accessible to content mining operations.

It's easy to see Google as a Web search and advertising company. But Dr. Halevy founded Nimble, which he sold to Actuate, a business intelligence company in 2002. Then he founded Transformic, Inc., which he sold to Google in 2006. I think it may be wise to practice a forest ranger's rule of thumb:



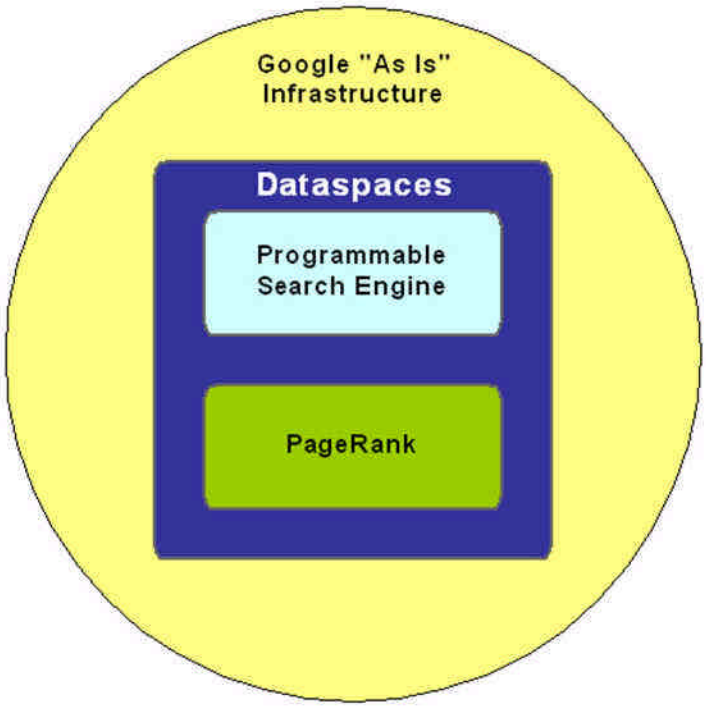
5.0 Google Patents

Google has indexed US patents. The system allows the user to view an abbreviated version of a patent, but the presentation is not likely to meet the needs of the professional researcher. You can test the system yourself. Just navigate to www.google.com/patents. You'll see the familiar Google interface. Enter a keyword, and the system responds with a list of results, relevance ranked according to Google's proprietary algorithm. Click on a result and you see a synopsis of the invention. The PDF of the original patent is available via a hot link to the often-unreliable USPTO servers.



6.0 Beyond an Index

Dr. Halevy's work, coupled with that of other Googlers and consultants to Google like Stanford's Dr. Widom, seems to point toward a federated approach to content. Google has been talking about "universal search" since May 2007. The one-off, grab bag called "Searchology" described a search function that puts different Google content in a single, relevance-ranked results list. You can see this type of function now if you navigate to Google.com in the US only at this time, and enter the query "Hillary Clinton". What makes this possible is the type of information integration and "dataspace" (not database) that Dr. Halevy's research enables. The diagram below shows a representation of the Google computing infrastructure, the original PageRank ranking service, the new Programmable Search Engine service developed by Ramanathan Guha, and Dr. Halevy's "dataspaces". The idea is that instead of a single list of results from Web content or Google Books, the results list would present many types of content, each relevant to the query, and instantly accessible from the Google interface.



7.0 An Enhanced Google “Stack”

If we zoom into this representation of the Google infrastructure (what I call the Googleplex), you will see that the top most layer is dataspaces. Beneath that are the PSE and the PageRank services. The foundation of this content processing engine is Google’s “as is” infrastructure. The notion of “as is” is quite important. Microsoft, Yahoo, and other competitors are developing their massively parallel, distributed, software as a service infrastructure. These companies are rushing and spending to convert their plans (their “to be” infrastructure) into reality. As I have documented in The Google Legacy and my more recent study, Google Version 2.0, Google is developing applications on its “stack”. Not surprisingly, Google is pushing into new areas such as content processing.

Competitors are trying to catch up to where Google was two or three years ago. This has profound implications for data and text mining.

8.0 The Future

This "stack" makes it possible for Google to slice and dice information. Its analytics capabilities reach far beyond the free Urchin Web analytics services. The company is ideally positioned to perform the types of analysis found in traditional data and text mining. Entit6y extraction, clustering, and relationships can be readily discerned in the dataspaces that Dr. Halevy has created first in the Nimble system and later in the Transformic system. This means that specialized data and text mining companies will find that Google offers some unique opportunities. First, the idea of spending to process certain open source or publicly-accessible content is an unnecessary cost for organizations now paying for this work. It may make more sense to use Google's indexes and its CSE or customizable search engine service. Second, Google's unique business model makes it possible for the company to process content that a single organization, regardless of its cash position, cannot afford to analyze; for example, Google Books aims to make searchable as many scientific, technical, and medical volumes as it can under law. In addition, Google wants to tackle technical journal content, Web logs, and rich media. Data mining and text mining that "drills down" into a narrow corpus will find that more relevant content may illuminate some difficult questions that today's systems cannot answer. For example, prior art may be easier to identify piggybacking on Google's processed content than trying to find the single, needed document using traditional research techniques. Finally, Google may elect to embrace partners as a means to competing in the data mining and text mining niche. If so, this means that today's dozens, maybe hundreds of competitors, may be faced with Google plus one or two preferred partners. Regard-

less of the future trajectory, Google is likely to exert significant gravitational force on data and test mining.

To answer the question, “Event horizon or sunset”, let me make three observations:

- Traditional approaches to data and text mining will be disrupted by Google’s content processing engine, its ability to deliver “universal search”, and Google’s wide use. No traditional content processing system can hope to match Google’s scale of operation and ubiquity.
- Google’s penchant for partnering and then favoring one or two high performers is likely to put increased financial pressure on many data and text mining concerns. These companies can survive as niche players, but the likelihood of a Thomson or Business Objects breaking out to become the dominant force in next generation text mining seems to be an improbable event. A glass ceiling on revenues will almost certainly change the funding patterns for innovation in this sector.
- Most traditional data and text mining vendors are generally unaware in the shift to enterprise publishing systems. Google’s moves in a yet more abstract and significantly larger sector are almost unknown. A lack of information about what’s happening in the information processing sector is not likely to lead to informed decisions. The most visible symptom of this problem may be rising prices, over-inflated claims, and confusing marketing messages.

Change will not be sudden. I plan on watching developments over the next nine to 24 months.

Stephen E. Arnold
Stiges, Spain
October 22, 2007