artesian™
Get to know

taking a sip from the firehose

An Artesian Whitepaper

artesiansolutions.com

It has been worked out that there are roughly 5,000 new pages of information being "added to the internet every single second of every day, related studies estimate that the number of characters or "bytes" of information currently accessible from the internet and corporate networks is fast approaching the ominous sounding 1 Yottabyte.

For those of us just getting used to terabyte hard drives this represents a trillion terabytes, i.e. 1024 bytes, which by anyone's standard is a very big number, perhaps bigger than a human can actually comprehend. Whatever the exact size is its fairly clear that something so big will be difficult to effectively exploit without some kind of organisation or structure. Bullet point number 7 in the corporate philosophy statement of Google says, "There is always more information out there", this is a heuristic that rings true for both companies and individuals, so how useful is the internet as a business resource? Our collective challenge is that a person can't read 40 billion WEB pages to extract the useful and relevant information, more importantly a person cannot re-read every page everyday just in case something new and important pops up, but luckily computers can.

Many companies have attempted to bring order to the chaos that is "the internet"; most notable since 1994 have been Lycos, Alta-vista, Google, Yahoo, MSN and Ask.com among others. The approach so far has mostly been to attack the problem from the perspective of "search", i.e. the consumer (business or person) have a question that is somehow captured in a single search phrase and then some software searches a pre-existing index of all the text to try and "find" matches. This approach works very well and is the corner stone of many killer applications such as Google search; however search is not the only kind of question that needs to be addressed.

Looking at the internet as a source of business intelligence.

For business people most questions are very difficult to boil down to a simple search phrase, a seemingly simple question like "what are my customers doing about green energy" is simply not something that search engines can answer. The reasons for this are numerous but probably the most obvious would be that search engines don't know anything about you the consumer; for example Google does not have a list of your customers. The other challenge that business has is that quite often the nature of business questions is much more temporal, for example its commonplace to ask about subjects like "sales" that are prefixed or suffixed with a temporal phrase for example "are sales increasing" meaning if I plot sales over time is the line going up? This is compared with a typical (personal) search problem like "where is the nearest Chinese restaurant to me now".

Time has long been a fundamental part of Business Intelligence (or BI for short) solutions, it would be unusual to see a business report that showed data which didn't have some kind of time dimension to it; for example costs compared to last year or projected revenue for next quarter etc. However business intelligence systems are almost exclusively inward looking, they work with structured sources of data like spreadsheets and databases created and maintained inside the company. What BI systems then do is aggregate and slice & dice that data by tangible dimensions such as products, years, people and places. Information on the internet however is usually much softer, less tangible; things like opinion, commentary and news etc. this kind of data tends to be much less accessible and full of ambiguity. However, it is evident that having visibility of how opinion regarding your service or product changes over time is potentially very valuable insight. The key question is how can we extract a tangible meaning from internet content such that we can apply business intelligence principals to it and exploit dimensional concepts to make that content navigable and measurable; in short we need to incorporate meaning and time.

Making WEB pages tangible by extracting their meaning

WEB pages (and other text based content) are notoriously difficult to extract meaning from; there is a whole branch of computer science devoted to exploring methods for doing it called Natural Language Processing or NLP for short. The challenge lies in the fact that human language is incredibly flexible and in almost every situation there are many, many ways of saying the same thing. For example, how could a typical news story communicate a simple concept like the stock price for company x has fallen?

Here are just a few examples:

- Stock crash for Company x
- Shares in company x fall 80% in a single day
- x bombs
- Equity bloodbath, company x suffers most

It is clear from this simple example that in order to understand the language used we need to also understand the context of the information and important parts of the culture that it came from. Human brains grow up within cultures and learn empirically over many years how to recognise context, this unspoken insight is natural for us, computers on the other hand don't "learn", they simply respond to pre-determined rules or "programs" that essentially need to cover every permutation and combination of context and meaning, something practically impossible when it comes to natural language understanding, there are just too many permutations and too few hard and fast rules to take this kind of approach.

The field of understanding content like WEB pages is at one of the cutting edges of computer science, enormous strides have been made over the past few years as the power available to practitioners has increased inversely to the cost of that power. We are now starting to see practical and useful techniques emerging from the labs that add huge value to information centred applications making it possible to treat unstructured content in similar deterministic ways as we do business intelligence information.

Surveillance as a compliment to search as a solution for business

At Artesian we believe that business cannot afford to ignore the internet as a source of business intelligence and that this source will continue to grow larger and more influential over time; however information derived from that source needs to include meaning and time in order to be truly useful. We believe it will be critical that unstructured content be transformed so that it can be structured by typical business dimensions such as prospects, clients, products and competitors etc. Armed with dimensional concepts it will be possible to extract value from that information by treating it like any other kind of business intelligence and subjecting it to measurement over time. We call this concept of applying meaning and tracking over time “surveillance” and we think that it is the missing link in terms of allowing organisations to effectively exploit the rich source of commercial insight that the internet represents. We believe that the early adopters of these kinds of “internet enhanced” applications will be the ones who not only help shape the ultimate mainstream solutions, but will also be the ones who benefit from the undoubted competitive advantage in being the first.

About the Author

Steve Borthwick, Chief Technology Officer.

As Chief Technology Officer has a 20 year track record in the software and ITservices industry and prior to joining Artesian, was Chief Technology Officer at MRM solution vendor Then-Solutions (acquired by Aprimo Inc in July 2004) pioneering the area of dynamic content management and publishing. As CTO of Whitelight Inc (acquired by Symphony Technology Group in January 2002) Steve led the development of Analytical Application Server technology including high end data aggregation and scenario modelling techniques. Steve has also held a series of senior positions at Cognos Inc, including Business Development Director for EMEA. Steve holds a degree in Computer Science.

About Artesian Solutions

Artesian Solutions – a provider of semantic-based sales intelligence and market intelligence technology delivered through the cloud which radically reduces the cost, time and pain of filtering the web to identify the key insights and sales triggers which drive commercial performance.

Artesian's clients include Barclays, SITA Suez, BusinessStream, Vodafone, JDA, EPICOR and many others.