Correlation Technology Solutions
Compared to
"Massive Semantic Infrastructure" Solutions

When enterprise or government face difficult, critical problems – intractable problems that simply must be dealt with – the solutions that are constructed in response to these problems are often "non-optimal". Typically, the solutions are expensive, and stunningly complex. Also typically, these solutions do not perform very well – despite their expense and complexity. These attributes – expense, complexity, poor outcomes – are especially prevalent in those cases where computer software is the basis for such solutions. These "non-optimal" software solutions can be found in many of the 1200 vertical market sectors for enterprise (identified by $NAICS_{[2007]}$), and in every sphere of government operations. Research from Make Sence Florida, Inc. has shown that non-optimal software solutions are likely to use one or more of three approaches:

"Massive Semantic Infrastructure Solutions" – Systems that require large natural language databases, ontologies, taxonomies, and concept repositories, and utilize tagging, threading, entity recognition, and other similar corpus analysis techniques in preparation for answering user queries.

"Subjective Statistical Model Solutions" – Systems that rely upon statistical models influenced by subjective human judgments in establishing base or conditional probabilities of events or outcomes – particularly those which purport to capture *all* possible events in a complex real-world domain. Such systems typically utilize Bayesian statistical techniques and include Neural Networks.

"Brute Force Computing Solutions" – Systems which achieve results from the power of modern day computers to perform a relatively simple process at high speed against large volumes of data. Keyword searches are a typical example.

The purpose of this document is limited to the examination of how "Massive Semantic Infrastructure Solutions" differ from Correlation Technology Solutions. A large well known enterprise software company which is referred to below as "Company A" and that company's primary product, which we call "M-Technology" is the example used in this discussion.

Company A is in fact our "poster-child" for what we call "non-optimal, massive semantic infrastructure solutions". We like to begin with the practical issues, because the practical aspects of a Company A solution illustrate perfectly why Company A's "M-Technology" compares so poorly to Correlation Technology.

We often like to recount this true story. At a NYC Search Engine Expo at which we presented in 2008, a senior staff member of a “Major US Government Financial Institution” stopped by our booth and, after listening to our explanation of Correlation Technology, started to complain about Company A - which his organization had purchased. He said, "for Company A to find a 21-word email I sent (in the past), I had to remember and enter into the search interface 20 of the words."

Here's why this happens. Before the Company A system can answer a single question, an enormous set of massive Natural Language databases must be installed and verified. Then, equally massive dictionaries, thesauri, "concept" repositories, ontologies, lexicons, and other semantic infrastructure components must be installed and linked. Then, the corpus (all of the documents) is subjected to indexing, threading, entity recognition, and other "associative" and "tagging" processes. These require days or weeks of dedicated server time and huge amounts of memory and data storage. Finally, the system is ready to do some work. But despite all of this effort, complexity and expense, the Company A system appears "stupid".

The Company A system appears "stupid" because Company A software is based entirely on an externally imposed "formal" construct of human language. The "meaning" part of "M-technology" in fact is constrained to those standard meanings and uses of words consistent with established academic models. Words are fixed in their allowed use as only specific parts of speech. The word proximities examined in texts are disregarded if they do not meet pre-set statistical thresholds of confidence. Syntactically modeled sentence decomposition is rigidly adhered to, and indexing schemes for "organic" keyword search are not much improved from their original implementations in the 1990's.

All of these formalisms are observed despite the fact that human expression is riotously, deliriously, chaotic and adaptive on a moment by moment basis. Writers of even the shortest communication incorporate cultural memes that no dictionary, no ontology, no concept map, no semantic infrastructure component could keep current or sort out. Humans create and utilize idiomatic, vernacular, and colloquial terms and uses for terms with astounding rapidity and ease, and with astounding confidence in the belief that such terms and every nuance of meaning carried by such terms will be perfectly understood and appreciated by the recipients of their expression (and they usually are). Trouble is, Company A (and its peers) can not make sense of anything not hardwired into the software's semantic components.

While it is certainly true that a corpus of only very formal documents – such as government reports, academic papers, and so on - will with the proper lexicons be well served by a Company A type approach, and while it is also true that Company A has obliged to provide facilities to users to "make their own lexicons" and to "define their own concepts" (so, with massive and amazingly time consuming and costly customization Company A’s product will work better), the fact remains that wherever human expression and comprehension is informal (such as the majority of email in an enterprise, human speech captured from transcripts, almost all the other categories of text

produced by amateur and professional writers for any purpose), Company A subjects its users to the possibility for the type of frustrations described above.

If the original text doesn't contain text which conforms to or is confined to the formal parameters of the academic models used, Company A can often have a lot of trouble in locating that text. In the last resort, a super-majority of "word matches" was required by Company A to find the employee's email, because all the "M-technology" was worthless. The same result could have been achieved with a universally available - and free - Unix text search utility.

Correlation Technology, in contrast, "permits" a far more "relaxed" and "natural" model of human language. Our one way, exhaustive transform of data into Knowledge Fragments (which we call "Acquisition") captures all the significant relations between words - as they are actually expressed in the text. Unlike "M-technology", Correlation Technology does not coerce the text into conformity with a set of formalisms or analyze the text using such formalisms. We "allow" every nuance to be captured without concern that some artificial rule is observed.

The Correlation process discovers knowledge from the corpus by constructing chains of iteratively associated Knowledge Fragments, and then analyzing the "Answer Space" (like the “result set” for RDBMS/SQL) of Correlations. Associations between words can be as formal or informal as desired or required for the application. We provide in the Correlation Technology Platform the ability to "dial in" more than 20 differing levels of "fuzzy association" that actually capture - without imposing any rules which prevent the discovery of knowledge - all the types of formalisms "understood" by "M-technology". Further, any additional "reference" preferred for associating words can be "plugged in".

By means of Correlation, knowledge is "emergent", meaning that the analysis of the Answer Space (a process we call "Refinement") will reveal the desired solutions - if they exist in the corpus. When the task is Enterprise Search, our Acquisition, Correlation and Refinement functions will reveal those emails, memos, or documents that the user wants.

Correlation Technology solutions are possible for every product offered by Company A. In each of these solutions, we believe the Correlation Technology approach will prove far more effective, far more flexible, and far more straightforward in implementation. While Correlation Technology solutions can be large scale, every Company A implementation dwarfs Correlation Technology implementations for an equivalent corpus. While the complexity of the Correlation Technology solution is obvious, that complexity does not flow from the hopeless attempt to capture in stone the torrent of human expression and comprehension, and in fact, Correlation Technology is intrinsically "simple".

For Business Inquiries:
Contact: Carl Wimmer
carl@makesence.us
Mobile: (702) 767-7001

For Technical Inquiries:
Contact: Mark Bobick
m.bobick@correlationconcepts.com
Mobile: (702) 882-5664