Cognitive Relevance Computation to enable Advanced Analytics and Predictive Engines

Joe Glick – August 18, 2012

Abstract

Science, business and government are struggling with the daunting challenges of transforming rapidly scaling stores of "big data" into evidence-based, actionable intelligence. The limited achievements and unrealized expectations recorded thus far by leading enterprises[1] underscore the obstacles to interdisciplinary knowledge integration, which is critical to discovery and value.

This paper presents a unified knowledge engineering paradigm wherein cognitive methods for relevance analysis define computational models for integrating, rationalizing and abstracting data into knowledge, enabling advanced analytics and predictive engines. Data points are modeled as a neural network of interacting systems, structures and scenarios, as well as their properties (behaviors and attributes). The scenarios drive relevance computation, and:

- are composed of semantic, procedural and/or episodic elements
- can be situated across the behaviors of systems and structures (scenarios share properties with the systems and structures)
- may be emergent properties of the systems and structures (scenarios share attributes with the systems and structures)

Scenario attributes include relational, hierarchical, unstructured and random data. Scenario elements compute relevance, and are abstracted as:

- Ontologies
- Taxonomies
- Categories (Chomsky)
- Agents (including external models)
- Rules

This approach is aligned with the objectives of the "Real-World Reasoning" initiative that DARPA launched in 2005. Having proven key modeling and abstraction methods in knowledge engineering projects at global organizations, we are currently working on an automated reasoning system combining "knowledge agents" and "learning agents" to accelerate discovery.

---

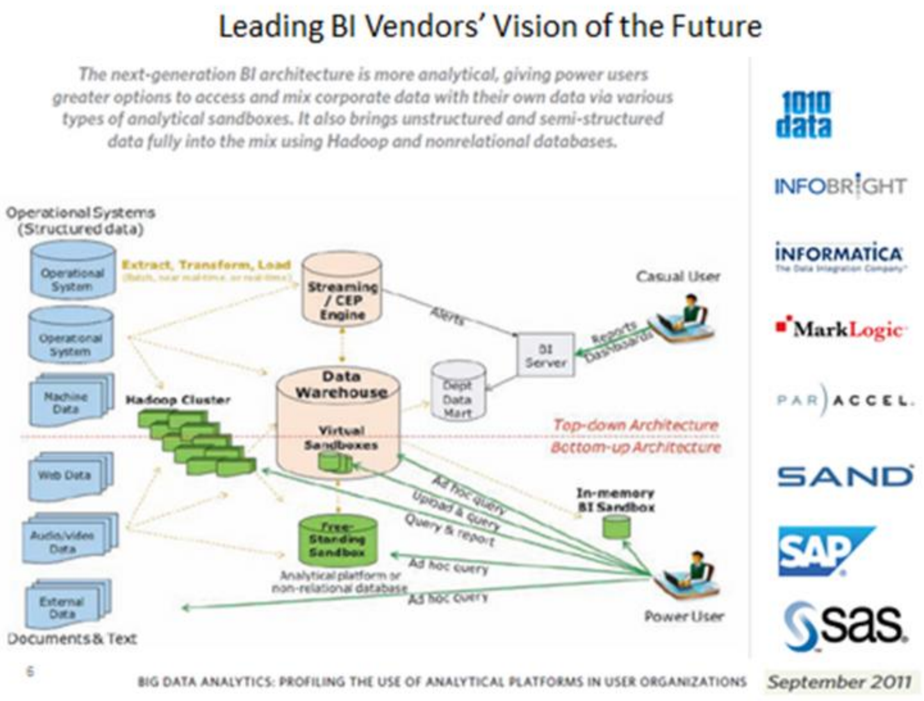
[1] http://www.zdnet.com/fords-big-data-chief-sees-massive-possibilities-but-the-tools-need-work-7000000322

The Obvious Challenges

Two of the key challenges to complex data integration – combinatorial complexity and system architecture silos – are well known, frequently discussed and driving the R&D efforts of major technology vendors.

The combinatorial complexity issue targeted by the DARPA initiative is driving IBM and five academic partners who jointly won the funding, to develop a neural chip[2], abandoning the traditional silicon chip design for a more biomimetic architecture. The goal is not just faster processing, but also the scaling of current capacity for combinatorial computation. Researchers at MIT have taken a different approach, seeking to mimic the brain's plasticity by modeling the activity of a single synapse using about 400 transistors[3]. While both efforts are significant from a research perspective, for those charged with enabling advanced analytics capabilities within their organizations these research initiatives serve only to emphasize the challenges ahead.

Business Intelligence (BI) vendors are seeking to leverage the "sandbox" data architecture approach utilized effectively by the developers of geographical positioning software, to enable complex, cross-platform query applications. The key strategy involves combining the sandbox architecture with in-memory analytics tools to break through the existing information barriers built over decades by traditional system architecture implementations, as seen below:



Leading BI Vendors' Vision of the Future

The next-generation BI architecture is more analytical, giving power users greater options to access and mix corporate data with their own data via various types of analytical sandboxes. It also brings unstructured and semi-structured data fully into the mix using Hadoop and nonrelational databases.

BIG DATA ANALYTICS: PROFILING THE USE OF ANALYTICAL PLATFORMS IN USER ORGANIZATIONS September 2011

---

[2] 8/24/2011 http://www.computerworld.com/s/article/print/9219288/IBM_brings_brain_power_to_exp...

[3] http://www.mit.edu/newsoffice/2011/brain-chip-1115.html

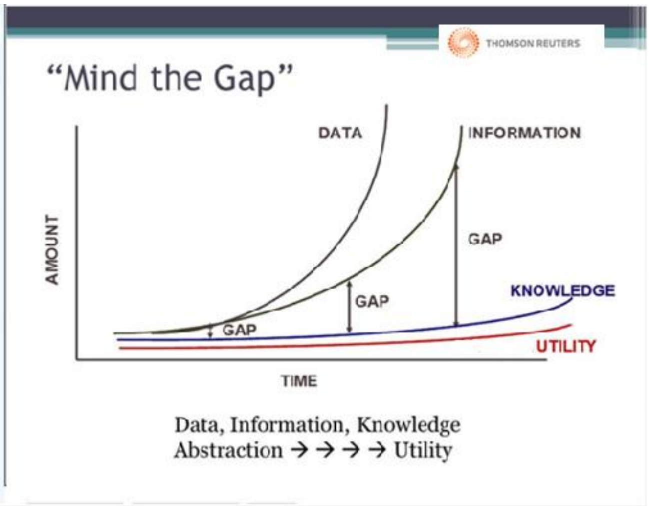
In the diagram above from a whitepaper funded by the listed vendors, the green components represent the future, but rather than depicting any new technology the architecture envisions new products and services being developed as an additional layer of infrastructure to enable queries across information silos. Additionally, the presented approach does not address the combinatorial complexity issue, which is the goal of the neural chip researchers.

The sandbox architecture is important to the relevance computing approach presented in this paper, but the intelligence needed to implement the cognitive methods that address scaling complexity requires neural sandboxes, and the utilization of that intelligence is beyond the capabilities of queries.

The significance of the challenges of combinatorial complexity and information silos is intuitive to analytics and knowledge engineering professionals, and it is likely that the level of associated research and product development will continue to increase. There exists another challenge, however, that is subtle and rarely discussed, but which is even more significant from the perspectives of technological innovation and value - utility.

Utility – the Subtle Challenge

According to a recent study by Symantec, digital information costs businesses 1.1 trillion dollars annually and enterprise information loads are projected to grow 67% this year[4]. Most of that money is spent on acquiring, storing, securing and maintaining the underlying data, as well as transforming the data into information and making it available to organizational users. The issue of utility – the extent to which the users are helped to accomplish their purpose – receives little attention and resources, as illustrated below:[5]



"Mind the Gap"

Data, Information, Knowledge
Abstraction → → → → Utility

---

[4] http://www.eweek.com/c/a/Data-Storage/Digital-Information-Costs-Businesses-11-Trillion-Annually-Symantec-825880/

[5] Center for Cancer Genomics and Computational Biology - Van Andel Research Institute – December 2011

The above study was performed in the field of medical research where utility is paramount, and the criticality of the issue as well as the limited progress made thus far are clearly illustrated. The diagram also identifies the gap between computed INFORMATION and KNOWLEGE as the primary problem domain.

Information is data in a useful structure, while knowledge is information in a useful CONTEXT. As demonstrated by the periodic table, the effectiveness of the transformations from information to knowledge is dependent on the adequacy and correctness of the ABSTRACTION methods, rather than the volume of available data or the speed of the search.

Mental Models define Utility

Since a human user needs to act on the output, the utility of the knowledge is dependent on the alignment of the output with the user's mental models, especially when the objectives are decision quality and learning. Cognitive research[6] has demonstrated that:

- the definition of the learning objective is not based solely on the accuracy of knowledge, but also on the subjectively and contextually determined utility of knowledge being acquired
- humans entertain multiple hypotheses and learn not only by modifying a single existing hypothesis but also by combining a set of hypotheses

The conclusions above are intuitive and are as applicable to decisions as to learning, but the issue of utility is frequently ignored and is a major cause of the too common disconnect between user needs and the functionality of delivered systems. Consequently, addressing the challenge of utility includes enabling the following technological capabilities:

- a contextual architecture that supports the definition and simultaneous interaction of multiple hypotheses and abstraction methods
- a relevance computation engine that can link the properties and attributes of the hypotheses to data across domains and levels of abstraction
- analysis and maintenance processes across the layers of context and hypothesis that are driven by a combination of information updates, computations and dynamic input from curators or users

This approach represents a radical departure from traditional knowledge management and analytics solutions. The approaches associated with current solutions continue to widen the gap between information and knowledge, making utility more elusive. To paraphrase Einstein, we cannot bridge the gap using the thinking and technology that is continuing to create it.

---

[6] Toshihiko Matsuka et al, Neurocomputing - August 2008

Relevance Analysis addresses the Challenges

Relevance analysis and computation addresses all three challenges – combinatorial complexity, information silos and utility, because it is the foundation of real-world reasoning. For decades, cognitive researchers have understood that flexible mental models created by contextual and subjective relevance processes enable the superior analytical and decision capabilities of experts in comparison with those of educated novices[7]. Additionally, recent research has shown that when asked to sort descriptions of real-world phenomena, novice students of the physical sciences sorted primarily by the domain, whereas experts sorted primarily by causal category[8], emphasizing that effective relevance analysis integrates procedural and episodic knowledge with the semantic approach to which the domain categorization is limited.

To align with this reality, knowledge engineering solutions need to be able to receive expert input in a way that captures their mental models of relevance within and across target domains, together with their contextual and subjective associations with processes and historic data. The captured mental models acting as agents in deterministic and stochastic interactions form the basis for real-world reasoning networks that can assist with the key challenges described.

Criticality of Relevance Computation

When the volume and complexity of the data puts exhaustive computation out of reach, the necessity of computing relevance is evident. However, relevance computation is not just a fallback position to cope with scaling data, but is usually the critical path to an effective solution. For example, the expression “real-time decision support” is appearing with increasing frequency in discussions of “big data” issues and objectives, but in the real world it is often an oxymoron. Real-time information is only actionable when it reports events for which all the following conditions exist:

1. the event is known and has been analyzed
2. a policy exists (i.e. a decision has already been made) for responding
3. a process exists to respond
4. resources have been allocated to respond

A current study found that delivering operational information to Mobile BI users reduced average decision time from 190 hours to 66 hours[9]. While the improvement is significant, the results show that after receiving the information, users still needed to make a significant investment of time and effort to achieve utility – i.e. make a decision, and we have no data on how frequently the users decided that the output did not really address the problem at hand.

---

[7] Mayer, R., 1992, *Thinking, Problem Solving, Cognition*, Freeman

[8] B. M. Rottman et al, 2011, Causal Systems Categories: Differences in Novice and Expert Categorizations of Causal Phenomena

[9] March 2012, Mobile BI 2012: Accelerating Business on the Move, Aberdeen Group

Relevance analysis is a core element of problem analysis, which is all about asking the right questions, so if a real-time solution is being considered, then the first question is: are we seeking to enable informed decisions or automate implementation of decisions already made? If large and complex data stores are involved, the solutions for decision support and event management are usually mission-critical and expensive propositions, and as a result relevance analysis is a critical success factor.

The impact of the relevance criticality is:

- the ***utility*** of information delivered for both discovery (research, event tracking, etc.) and analysis (decision support, planning, etc.) is directly proportional to the *granularity and precision of the context architecture*
- the ***total cost of ownership (TCO)*** is inversely proportional to the *human expertise that is embedded* in the relevance analysis processes

Consequently, to the extent that context can be precisely architected and relevance correctly computed, it increases the value and utility of the outputs and reduces the TCO. This is possible because problem analysis becomes an embedded driving force in operational and maintenance processes, not just a high-level exercise at the beginning of a project. System architects are well aware of the cost-of-errors heuristic: errors not detected at the requirements stage cost ten times as much to repair at the design and construction stage, and ten times more at the testing stage. Defining requirements to solve the wrong problems can add orders of magnitude to this painful reality, but relevance analysis is the critical path to ensure that the solution assembles the correct data and methods to transform information into knowledge, and knowledge into utility.

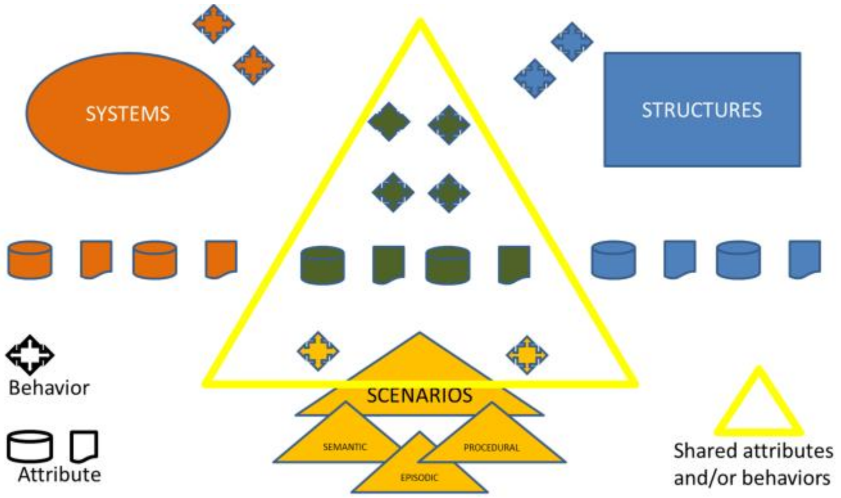
The Expertool Relevance Analysis Paradigm

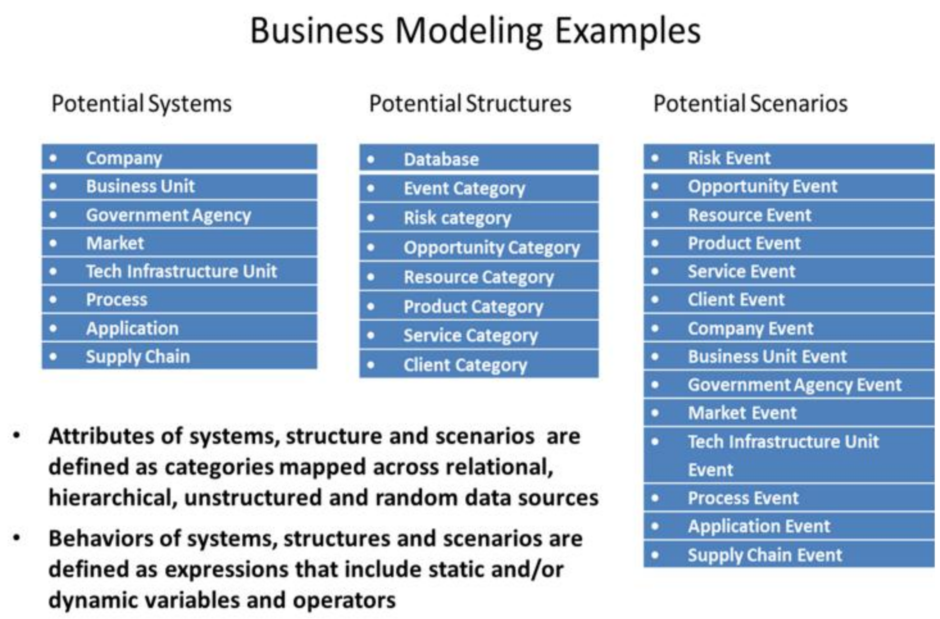
The Expertool relevance analysis and computation methodology is a unified knowledge engineering paradigm wherein cognitive methods for relevance analysis define computational models for integrating, rationalizing and abstracting data into knowledge, enabling advanced analytics and predictive engines. Data points are modeled as a neural network of interacting systems, structures and scenarios, as well as their properties (behaviors and attributes). The scenarios drive relevance computation, and:

- are composed of semantic, procedural and/or episodic elements
- can be situated across the behaviors of systems and structures (scenarios share properties with the systems and structures)
- may be emergent properties of the systems and structures (scenarios share attributes with the systems and structures)

This real-world reasoning approach enables the construction of models that integrate highly diverse elements and information sources to enable exploration and discovery to a scope that traditional information architecture cannot accommodate.

The holistic environment can be illustrated graphically as follows:





Business Modeling Examples

Potential Systems

- Company
- Business Unit
- Government Agency
- Market
- Tech Infrastructure Unit
- Process
- Application
- Supply Chain

Potential Structures

- Database
- Event Category
- Risk category
- Opportunity Category
- Resource Category
- Product Category
- Service Category
- Client Category

Potential Scenarios

- Risk Event
- Opportunity Event
- Resource Event
- Product Event
- Service Event
- Client Event
- Company Event
- Business Unit Event
- Government Agency Event
- Market Event
- Tech Infrastructure Unit Event
- Process Event
- Application Event
- Supply Chain Event

- **Attributes of systems, structure and scenarios are defined as categories mapped across relational, hierarchical, unstructured and random data sources**
- **Behaviors of systems, structures and scenarios are defined as expressions that include static and/or dynamic variables and operators**

The above described approach represents a significant enhancement to current information systems and infrastructures. To help envision the possibilities, Expertool Paradigm, LLC offers its unique "Solve a Problem" software evaluation program which includes training for the evaluator(s) to use the platform to address a current information challenge with which the organization is struggling. The evaluation can be requested at www.expertool.com.

Author contact: joeglick@expertool.com