Published on December 7, 2005

The Data Center

Engineering Marketing Science: Semantic Search

Edmund W. Schuster and David L. Brock

The Data Center, Massachusetts Institute of Technology, Building 35, Room 234, Cambridge, MA 02139-4307, USA

ABSTRACT

Businesses place a great deal of emphasis on cost reduction. However, with increasing inflation, the cost reduction strategy is having less impact on enhancing profits. Many companies are now focusing on operational ways to increase revenue in addition to cost reduction. This article examines the idea of using semantic-based methods to enhance the precision of Internet searchers. Achieving greater precision in search has significant implications for Internet advertising, a proven means for increasing revenue.

ABOUT THE AUTHORS

Edmund W. Schuster has held the appointment of Director, Affiliates Program in Logistics at the MIT Center for Transportation and Logistics and is currently working at The Data Center as Co-Director - Administration and researcher. His interests are the application of models to logistical and planning problems experienced in industry. He has a bachelor of science from The Ohio State University and a master in public administration from Gannon University with an emphasis in management science. Ed also attended the executive development program for physical distribution managers at the University of Tennessee and holds several professional certifications. Ed can be reached at Edmund_w@mit.edu

David L. Brock is Principal Research Scientist at the Massachusetts Institute of Technology, and co-founder and a Director at the Auto-ID Center (now EPCGlobal, Inc. and Auto-ID Laboratories). The Center was an international research consortium formed as a partnership among more than 100 global companies and five leading research universities. David is also Assistant Research Professor of Surgery at Tufts University Medical School and Founder and Chief Technology Officer of endoVia Medical, Inc., a manufacturer of computer controlled medical devices. Dr. Brock holds bachelors' degrees in theoretical mathematics and mechanical engineering, as well as master and Ph.D. Degrees, from MIT. Dave can be reached at dlb@mit.edu

1.0 INTRODUCTION

Cost cutting is always a good strategy for business. However, there are limits to the rate of profit growth when the sole focus is gaining operational efficiencies. With inflationary pressures, the ability to reduce costs is becoming increasingly difficult for manufacturing and service firms alike.

The future is enhanced revenue generation through sophisticated application of marketing science, engineering technology, and supply chain management. Integrating these three disciplines is the path leading to revenue growth.

Addressing this goal, the MIT Data Center is a new initiative charged with researching, developing, and prototyping the languages, protocols, and technologies to integrate data and models across global networks, creating interoperability on a scale not currently possible. The technologies and standards created by the Center will be open and freely distributed. Ultimately, this work lays the foundation for a new *Intelligent Information Network*.

The centerpiece of the efforts of the Center is the M Language. Designed as an open system, M utilizes a central dictionary based on the Wiki concept, which is an innovative way to achieve consensus concerning the definition of a word. With the Wiki approach, M overcomes a significant amount of semantic ambiguity currently associated with Internet search engines and standards.

The MIT Data Center concentrates on the parallel development of computer infrastructure along with real-world applications. It is organized by industry verticals such as the process industries (consumer goods, food, pharmaceutical, agricultural, and retailing), heavy industries (petroleum and automobile), heath care (medicine and biotechnology), and governmental operations. To date, there are five specific areas of focus:

1. Building an intelligent **network that links models to data**. The goal is to improve the speed of modeling in practice, encourage the re-use of model elements, and find new combinations of models.

2. Industry has expressed a great interest in the idea of **translating data at the edge of computing systems**. This opens opportunities to create new data sets from a number of different sources. One example involves the integration of weather data with demand data. This would be useful for various marketing science studies and for logistics and supply chain management. There has also been interest in building an interoperable data system for conglomerates, where each operating unit has its own approach to organizing data.

3. Another area is the development of an **Internet Search tool that uses the definition of the word**, rather than the word itself, to conduct a search of web pages and perhaps XML code. Accomplishing this capability for even a small portion of the Internet will increase by an order of magnitude the precision of search. This has potential to increase the value of online advertising, already a rapidly growing market.

4. Various forms of **visualization** of data through a tangible user interface.

5. A final area involves using the M language to improve **data quality**.

Though all five of these areas have the potential to improve efficiency within organizations, the area of internet search based on a specific definition of a word or noun phrase will have perhaps the greatest impact on revenue growth. Definition-based searches even for a small portion of the Internet will improve the precision of searches. With the M Language, advertising companies will be able to target specific customers as part of the search process. This will improve the effectiveness of advertising in increasing revenue.

The remainder of this article examines the idea of creating browsers that utilize definition-based search.

2.0 THE NEXT INNOVATION IN INTERNET SEARCH

At the most basic level of communication, words are the glue that connects nearly everything together. The power of words can give descriptive meaning to the most complex physical objects existing in business or nature, and to the most diffuse ideas that exist only in the mind. Data is described with words, information is word based, and computer code uses words as a way of communicating the various operators available to programmers. Words establish not only the limits of human imagination and intellect, but also the possibilities for the computing systems of the future.

Every word has at least one definition and when used in conjunction with other words there are a near infinite number of sentences possible to create. For example, a simple cartoon shown to twenty-five different people will generate twenty-five individually unique verbal perspectives if each writes a single sentence about what they see (Lederer, 1991, p. 16). Dictionaries organize and define words used to make sentences, and a weak ontology based on groupings such as synonyms and antonyms provides the relationships that the human mind can fathom into linguistic communication.

In spite of the almost limitless capacity of human language, describing physical or abstract objects in a consistent machine understandable way remains a challenging objective for both practitioners within business as well as for computer scientists. Although English is a powerful tool for communicating meaning through words, noun phrases, and sentences of varying patterns and complexity, this ability is under-utilized by modern computing systems.

The fundamental problem with employing words as a descriptor is that a single word can have several different definitions and multiple words can have the same definition. This paradox means that natural language often does not have the internal consistency required for strait-forward application as an identifier or a unit of meaning within computer systems.

Complicating matters, the intricacy of meaning increases dramatically when dealing with the noun phrases and sentences needed to describe abstractions. Given this property of English, it is impossible with current technology to conduct a semantically precise, computer-based search of information contained in web pages (HTML), quantitative data tagged with words, news feeds comprised of text files, complex mathematical models, or any other situation where words describe physical or abstract objects.

To move beyond the crude level of semantic search that exists today requires the incorporation of specific word definitions as a criterion of search. Since common words like "table" can have several definitions (i.e. furniture or array of numbers), it is imperative that a search be conducted on a single definition rather than the word alone. With a definition-based search, the number of matches will significantly decrease, resulting in much more precision. This is an important goal for firms such as Google that depend on the precision of searches to reach the demographic groups advertisers wish to target.

Though industry has not ignored the problem of definition-based search, the approaches employed have fallen short of transforming the Internet from a sea of unstructured data and information to an organized body of resources where exact semantic searches are a reality.

3.0 THE STANDARD APPROACH

The current solution to the problem of definition-based searches involves relying upon various standards groups that spend a great deal of time attempting to derive a single, universal definition for a commonly used word or noun phrase that fits all contexts within a particular industry. Several groups have gone as far as to employ the practice of creating camel case words, a situation where new words are formed by combining two or more existing words. The new camel case word then takes on a single meaning.

This practice expands the number of words used to describe objects; however, various industries might still assign different definitions to a single camel case word. In addition, a loss of flexibility occurs when combining strings of words into new words. Some examples of standards groups include RosettaNet, the National Retail Federation (NRF), Association for Retail Technology Standards (ARTS), the Uniform Code Council (UCC) Global Data Dictionary and many others. Though not participating in formal standards development, nearly all non-profit professional groups establish a dictionary to give common meaning for frequently used words. Several consortia like RossettaNet go a step further in developing standard terminologies and in some cases use XML-based schema as a means of describing business processes and data, and other abstract objects commonly used in commerce.

Though the single definition method has worked well in highly structured situations, each industry segment tends to become unique regarding the words and definitions employed. This insular approach increases transactional efficiency within a particular industry. However, industry specific definitions also sacrifice opportunities to share data, models, and other abstractions across industries because of the lack of a universally agreed upon definition that is globally visible. With this approach, any measure of Internet-wide semantic search based on word definitions becomes impossible to accomplish in practice. Typically, industry consortia do not provide the proper keys necessary to assign different definitions to the same word. This is a further hindrance to building system-wide Internet search capabilities based on word definition.

4.0 LIMITS OF THE STANDARDS APPROACH

Though developing a single definition for words or noun phrases works well within science and engineering, there are significant limitations when applying this approach to business situations that involve the description of a physical or abstract object. Researchers have employed various computer-based techniques, such as Artificial Intelligence (AI), in an effort to solve the problem of interpreting the meaning of language associated with an object.

These approaches have largely failed in practice because the meaning of a word, a phrase, or a sentence used to describe an object depends on the semantics of each word, the syntax or order of usage, and the context in which the word(s) appear. The unique properties of the human mind can determine contextual relevance, and then figure out how several relevant variables are associated (Deacon, 1997, p. 48). To date, machine languages have not duplicated this human property. Further, most AI techniques rely upon deductive reasoning where general concepts are applied to solve specific situations in terms of meaning. This becomes difficult to do in practice, because most meaning is inductive with a strong dependence on specific context.

5.0 INTEROPERABILITY IN PRACTICE

The current inability to interoperate and search data, information, and models arises from the need for translation between various independent dictionaries (situated in numerous standards groups), which involves writing a custom computer program for each translation. This situation becomes even more complex as the numerous dictionaries currently in existence undergo revision. Some estimate that for every n words contained in existing dictionaries there needs to be at least n2 translations as part of normal communication within industry. As the number of words used to describe objects in a machine understandable way expands, and dictionaries become larger, the volume of translations will become untenable, thus making inter-industry data sharing impossible to achieve in practice.

In the area of computer transaction systems, single-term standardization has its roots in several hundred years of engineering standards development, where the historic goal was precision in definition for highly specific situations along with universal adoption of the

agreed upon definition. In the case of commerce, most of these standards efforts have focused on words used in transactions with little emphasis placed on describing complex objects such as data, information, business processes, and mathematical models in a common, interoperable way. There are certainly many opportunities in business to expand the range of objects eligible for machine understandable description and semantic search based on definition.

6.0 ESTABLISHING SEMANTIC SEARCH

The essence of developing future machine understandable semantics depends on the fact that language, and English specifically, is relative in meaning based on the intended usage and context by those who originate the communication or search. This becomes apparent when comparing the definition of words used in different business and academic disciplines.

Since all words have definitions residing in various dictionaries, and all words are subject to classification, it is possible to design a computer system that utilizes multiple definitions for a single word, yet maintains system-wide consistency in relative meaning. Though complete machine contextual understanding of sentences is a distant goal, current computer technology, given the proper architecture, is capable of applying relative meaning to words and noun phrases used as descriptors for abstract objects like data, information, and mathematical models.

Relying on words and noun phrases as descriptors with multiple meanings defined in a centralized open dictionary, drastically improves the prospects for finding an exact semantic match when conducting an Internet-wide search based on a single definition. This offers the possibility of searching and matching data, information, and models from different disciplines, creating new types of interaction that do not currently exist because of barriers in description, definition, or format.

The work of the MIT Data Center involves researching and developing new computer architectures that utilize semantics to enhance search, along with improved ways of linking semantically labeled elements, either data or mathematical models, together within a network that spans the Internet. By using a single set of standards for describing abstractions such as data, information, and models, all with relative meaning, the M Language and Dictionary currently under development at the Center serves as a base to establish future semantic interoperability.

Building upon the existing standards of the W3C, including XML, M is an open system with a global dictionary based on the Wiki concept. This enables translation and interoperability of data and models at the edge of computing networks, using common definitions for words and noun phrases that are resident in the global Wiki dictionary. With this computer architecture, the process of semantic modeling can take place where mathematical models and data are described in a semantically precise way that allows for search, combination, and re-use of model elements and data. Using this approach, data and models become atomic with precise semantic definitions expressed as either individual words or noun phrases that are machine understandable.

7.0 INTEGRATION AND SEMANTICS

As Professor Grosof of MIT once commented during a public lecture, there are only several certainties in the world, "death, taxes, and integration (Grosof, 2005)." Interoperability through greater semantic integration will be one of the major IT trends of the next ten years. Software firms such as SAP already have major programs in place to improve integration and interoperability by using Web Services as a means of delivering ERP software to medium sized firms. The current movement toward open systems, recently highlighted through business acquisitions by IBM, will further enhance the need for Internet-wide semantic capabilities.

Even before the advent of computer based communication and networks, scholars from many different disciplines have often noted the importance of language and the value of a universal dictionary to the overall health of nations and to the advantage of commerce. Samuel Johnson, the creator of the first English dictionary, once commented, "Languages are the pedigrees of nations (Lederer 1991, p. 102)."

In the future, computer languages with a shared global dictionary with become the pedigrees of businesses. An important property that will establish the pedigree of a computer language and dictionary will be the ability to perform a semantic search based on one of multiple definitions for a word or noun phrase. These definitions will be historic in nature, capturing the changes in meaning that occur through time. To this end, the MIT Data Center is actively pursuing the research, development, and application activities needed to build the next generation of computer systems that will connect physical and abstract objects through words and noun phrases of different definitions. The future of computing will depend on new types of semantic-based systems capable of organizing and searching the ever-increasing volumes of data currently generated by business.

REFERENCES

Deacon, Terrence, W., *The Symbolic Species*, New York: W.W. Norton, 1997.

Grosof, Benjamin, "Lecture on Semantic Web and Semantic Web Services," MIT, April 5, 2005.

Lederer, Richard, *The Miracle of Language*, New York: Simon & Schuster, 1991.

NOTES