Discovery Informatics

William W. Agresti
Johns Hopkins University, USA

INTRODUCTION

Discovery informatics is an emerging methodology that brings together several threads of research and practice aimed at making sense out of massive data sources. It is defined as "the study and practice of employing the full spectrum of computing and analytical science and technology to the singular pursuit of discovering new information by identifying and validating patterns in data" (Agresti, 2003).

BACKGROUND

In this broad-based conceptualization, discovery informatics may be seen as taking shape by drawing on more of the following established disciplines:

- **Database Management:** Data models, data analysis, data structures, data management, federation of databases, data warehouses, database management systems.
- **Pattern Recognition:** Statistical processes, classifier design, image data analysis, similarity measures, feature extraction, fuzzy sets, clustering algorithms.
- **Information Storage and Retrieval:** Indexing, content analysis, abstracting, summarization, electronic content management, search algorithms, query formulation, information filtering, relevance and recall, storage networks, storage technology.
- **Knowledge Management:** Knowledge sharing, knowledge bases, tacit and explicit knowledge, relationship management, content structuring, knowledge portals, collaboration support systems.
- **Artificial Intelligence:** Learning, concept formation, neural nets, knowledge acquisition, intelligent systems, inference systems, Bayesian methods, decision support systems, problem solving, intelligent agents, text analysis, natural language processing.

What distinguishes discovery informatics is that it brings coherence across dimensions of technologies and domains to focus on discovery. It recognizes and builds upon excellent programs of research and practice in individual disciplines and application areas. It looks selectively across these boundaries to find anything (e.g., ideas, tools, strategies, and heuristics) that will help with the critical task of discovering new information.

To help characterize discovery informatics, it may be useful to see if there are any roughly analogous developments elsewhere. Two examples—knowledge management and core competence—may be instructive as reference points.

Knowledge management, which began its evolution in the early 1990s, is the practice of transforming the intellectual assets of an organization into business value (Agresti, 2000). Of course, before 1990, organizations, to varying degrees, knew that the successful delivery of products and services depended on the collective knowledge of employees. However, KM challenged organizations to focus on knowledge and recognize its key role in their success. They found value in addressing questions such as the following:

- What is the critical knowledge that should be managed?
- Where is the critical knowledge?
- How does knowledge get into products and services?

When C. K. Prahalad and Gary Hamel published their highly influential paper, "The Core Competence of the Corporation" (Prahalad & Hamel, 1990), companies had some capacity to identify what they were good at. However, as with KM, most organizations did not appreciate how identifying and cultivating core competences (CC) may make the difference between being competitive or not. A core competence is not the same as what you are good at or being more vertically integrated. It takes dedication, skill, and leadership to effectively identify, cultivate, and deploy core competences for organizational success.

Both KM and CC illustrate the potential value of taking on a specific perspective. By doing so, an organization will embark on a worthwhile reexamination of familiar topics—its customers, markets, knowledge sources, competitive environment, operations, and success criteria. The claim of this article is that discovery informatics represents a distinct perspective, one that is potentially highly beneficial, because, like KM and CC, it strikes at what is often an essential element for success and progress—discovery.

MAIN THRUST

Both the technology and application dimensions will be explored to help clarify the meaning of discovery informatics.

Discovery Across Technologies

The technology dimension is considered broadly to include automated hardware and software systems, theories, algorithms, architectures, techniques, methods, and practices. Included here are familiar elements associated with data mining and knowledge discovery, such as clustering, link analysis, rule induction, machine learning, neural networks, evolutionary computation, genetic algorithms, and instance-based learning (Wang, 2003).

However, the discovery informatics viewpoint goes further, to activities and advances that are associated with other areas but should be seen as having a role in discovery. Some of these activities, like searching or knowledge sharing, are well known from everyday experiences.

Conducting searches on the Internet is a common practice that needs to be recognized as part of a thread of information retrieval. Because it is practiced essentially by all Internet users and involves keyword search, there is a tendency to minimize its importance. Search technology is extremely sophisticated (Baeza-Yates & Ribiero-Neto, 1999). People always have some starting point for their searches. Often, it is not a keyword, but a concept. So people are forced to perform the transformation from a notional concept of what is desired to a list of one or more keywords. The net effect can be the familiar many-thousand hits from the search engine. Even though the responses are ranked for relevance (a rich and research-worthy subject itself), people may still find that the returned items do not match their intended concepts.

Offering hope for improved search are advances in concept-based search (Houston & Chen, 2004), more intuitiveness to a person's sense of "find me content like this," where *this* can be a concept embodied in an entire document or series of documents. For example, a person may be interested in learning which parts of a new process guideline are being used in practice in the pharmaceutical industry. Trying to obtain that information through keyword searches typically would involve trial and error on various combinations of keywords. What the person would like to do is to point a search tool to an entire folder of multimedia electronic content and ask the tool to effectively integrate over the folder contents and then discover new items that are similar. Current technology can support this ability to associate a fingerprint with a document (Heintze, 2004) in order to characterize its meaning, thereby enabling concept-based searching. Discovery informatics recognize that advances in search and retrieval enhance the discovery process.

This same semantic analysis can be exploited in other settings, such as within organizations. It is possible now to have your e-mail system prompt you, based on the content of messages you compose. When you click *send*, the e-mail system may open a dialogue box (e.g., Do you also want to send that to Mary?). The system has analyzed the content of your message, determining that, for messages in the past having similar content, you also have sent them to Mary. So the system is now asking you if you have perhaps forgotten to include her. While this feature can certainly be intrusive and bothersome unless it is wanted, the point is that the same semantic analysis advances are at work here as with the Internet search example.

The informatics part of discovery informatics also conveys the breadth of science and technology needed to support discovery. There are commercially available computer systems and special-purpose software dedicated to knowledge discovery (see listings at http://www.kdnuggets.com/). The informatics support includes comprehensive hardware-software discovery platforms as well as advances in algorithms and data structures, which are core subjects of computer science. The latest developments in data sharing, application integration, and human-computer interfaces are used extensively in the automated support of discovery. Particularly valuable, because of the voluminous data and complex relationships, are advances in visualization (Marakas, 2003). Commercial visualization packages are used widely to display patterns and to enable expert interaction and manipulation of the visualized relationships.

Discovery Across Domains

Discovery informatics encourages a view that spans application domains. Over the past decade, the term has been associated most often with drug discovery in the pharmaceutical industry, mining biological data. The financial industry also was known for employing talented programmers to write highly sophisticated mathematical algorithms for analyzing stock trading data, seeking to discover patterns that could be exploited for financial gain. Retailers were prominent in developing large data warehouses that enabled mining across inventory, transaction, supplier, marketing, and demographic databases. The situation (marked by drug discovery informatics, financial discovery informatics, etc.) was