Discovery Informatics from Data to Knowledge

William W. Agresti
Johns Hopkins University, USA

INTRODUCTION

It is routine to hear and read about the information explosion, how we are all overwhelmed with data and information. Is it progress when our search tools report that our query resulted in 300,000 hits? Or, are we still left to wonder where is the information that we really wanted? How far down the list must we go to find it?

Discovery informatics is a distinctly 21st century 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

The rapid rise in the amount of information generated each year may be quite understandable. After all, the world's population is growing, and countries like China and India, with very large populations, are becoming increasingly influential worldwide. However, the real reason why people are confronted with so much more information in their lives and work is that the information has real benefits for them. However, these benefits are not always or often realized, and therein lay the motivation for discovery informatics.

Companies today are data mining with more highly granular data to better understand their customers' buying habits. As a result, there is pressure on all businesses to attain the same level of understanding or be left behind – and being left behind in the 21st century can mean going out of business. Not-for-profits are becoming equally adept at mining data to discover which likely donors are most cost-effective to cultivate. Increasing granularity enables more targeted marketing, but with more data requiring more analysis. A co-conspirator in this infoglut is the declining cost to store the data. Organizations don't need to make choices on what data to keep. They can keep it all.

The task of making sense out of this burgeoning mass of data is growing more difficult every day. Effectively transforming this data into usable knowledge is the challenge of discovery informatics. In this broad-based conceptualization, discovery informatics may be seen as taking shape by drawing on more established disciplines:

- **Data analysis and visualization:** analytic frameworks, interactive data manipulation tools, visualization environments
- **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:

- 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 and 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 competencies (CC) may make the difference between 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 re-examination of familiar topics: its customers, markets, knowledge sources, competitive environment, operations, and success criteria. The claim of this chapter 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.

Embracing the concept of strategic intelligence for an organization, Liebowitz (2006) has explored the relationships and synergies among knowledge management, business intelligence, and competitive intelligence.

MAIN THRUST OF THE CHAPTER

This section will discuss the common elements of discovery informatics and how it encompasses both the technology and application dimensions.

Common Elements of Discovery Informatics

The only constant in discovery informatics is data and an interacting entity with an interest in discovering new information from it. What varies, and has an enormous effect on the ease of discovering new information, is everything else, notably:

- Data:
  - Volume: How much?
  - Accessibility: Ease of access and analysis?
  - Quality: How clean and complete is it? Can it be trusted as accurate?
  - Uniformity: How homogeneous is the data? Is it in multiple forms, structures, formats, and locations?
  - Medium: Text, numbers, audio, video, image, electronic or magnetic signals or emanations?
  - Structure of the data: Formatted rigidly, partially, or not at all? If text, does it adhere to known language?
- Interacting Entity:
  - Nature: Is it a person or intelligent agent?
  - Need, question, or motivation of the user: What is prompting a person to examine this data? How sharply defined is the question or need? What expectations exist about what might be found? If the motivation is to find