CDI-Type II: A Knowledge Web: Turning Raw Data into Accessible Knowledge

David W. Embley, Brigham Young University

The current web is a web of linked pages. Frustrated users search for facts by guessing which keywords or keyword phrases might lead them to pages where they can find facts. Can we make it possible for users to search directly for facts? Equivalently, can we turn the web into a web of facts (instead of a web of pages containing facts)? Ultimately, can the web be a knowledgebase that can provide direct answers to factual questions?

The answer to these questions calls for a move from data on the web to knowledge on the web—a move to distilling knowledge from the wealth of heterogeneous digital data. But how? Our computational alembic must turn raw symbols contained in web pages (or other source documents) into knowledge and make this knowledge effortlessly accessible via the web. We face three real challenges: (1) automatic (or near automatic) creation of extraction ontologies, (2) automatic (or near automatic) annotation of web pages with respect to these ontologies, and (3) simple, but accurate, query specification, usable without specialized training. Meeting these basic challenges will simplify knowledge-web content creation and access to the point that the grand vision of a web of knowledge will become a reality.

Based on—(1) current work that shows how to automatically/semi-automatically reverse-engineer collections of tables and semi-structured documents into extraction ontologies (NSF Grants 0414644 & 0414854); (2) prior work that shows how to automatically annotate web documents with extraction ontologies (NSF Grant 0083127); and (3) current and prior work that shows how to match free-form queries with extraction ontologies in order to generate structured queries over populated ontologies (NSF Grants 0083127, 0414644, & 0414854)—we propose to innovatively combine and extend all this work to meet these challenges. The end result is a scalable architecture for adding a knowledge-content layer over information-rich pages on the web that allows users to query for facts using free-form constraints, instance values, and concept keywords. Moreover, with provenance references linked directly to fact-origination statements, users can check retrieved facts in original source documents. The extent of required preprocessing, the accuracy of extracting facts from the web, and the range of user accessibility will be quantitatively evaluated over governmental, scientific, retail, and historical data.

Intellectual Merit. The research work: (1) provides an answer to the question about how to turn syntactic symbols into semantic knowledge; (2) shows how to create a web of knowledge; (3) shows how to establish a workbench with toolkits to convert heterogeneous digital data into knowledge under the auspices of an ontology; (4) explores the synergistic interplay among ontology, epistemology, and logic for the advancement of knowledge, providing new ways to think computationally about what knowledge is and how knowledge is acquired; and (5) provides a way for untrained users to query and reason over fact-filled ontologies.

Broader Impact. The research work has the potential to help people: (1) harvest and make available facts from the wealth of available heterogeneous digital data; (2) harness and manage community knowledge with the objective of enhancing human cognition; (3) make facts on the web (as well as pages) easily searchable by the general public; (4) provide a practical set of tools for knowledge management; and (5) involve knowledge workers from various disciplines in a community-wide effort to convert data into knowledge.

Educational Influence. Besides benefiting the general public, we also intend to promote training and cross-fertilization between departments and universities. Our research team is housed in four departments (Computer Science, Information Systems, Linguistics, and Electrical, Computer, and Systems Engineering) at two institutions (BYU and RPI). We will actively seek participation of underrepresented groups and demographic diversity. Our current research groups include female students (4) and students from the US (7) and Asia (4). To complement our cadre of CS research students and to bolster our knowledge-content areas, we will recruit an informal advisory committee made up of students majoring in government, biology, marketing, and history.