DATA EXTRACTION

CHALLENGE

People want to know! And so do government agencies, information providers, search-and-retrieval companies, electronic publishers, corporate enterprises, and business-intelligence professionals. But they're swamped with volumes of unstructured data spewed forth from search engines, corporate intranets, news feeds, and the increasing global Internet. They want critical information extracted automatically at the level of a human expert.

Critical information is difficult to locate. Once located, its incompatible formats make it difficult to use effectively. Large volumes of unstructured text must be digested into an easy-to-use, organized, uniform format to support querying, focused searching, personalized information products, and on-line transaction systems.

The challenges are: automatic search, automatic extraction, automatic integration, automatic analysis, and automatic summarization so—that people can concentrate on tasks that require human intelligence.

VISION

We propose to address these challenges by applying research/technology ideas in conceptual modeling and ontologies. A conceptual-model-based ontology provides a mechanism to represent knowledge, store information, and give symbols a specific meaning in a particular context. These ontologies can be leveraged to guide, combine, and interpret raw units of information and provide the basis for high-quality search and information extraction, integration, analysis, and summarization.

We do not mean to repeat the creation of top-level ontologies of broad knowledge domains. Instead, we envision small, application-specific ontologies. Our thesis is this: finding, extracting, structuring, and synthesizing information is easier given a conceptual- model-based ontology.

RESEARCH PROBLEMS

So, how do we build conceptual-model-based application ontologies to achieve these objectives? How do we algorithmically: elicit keywords and phrases from ontologies (so search engines can locate potential sources of information); break up sources of information to match data records in an ontology; extract semantic information from information sources; use natural-language and ontological information to match data records; and extract such data widgets as people, places, zipcodes, currencies, and time?

How do we: add a measure of certainty to extracted information; query, analyze, integrate, and summarize information; evaluate different displays of condensed information—easy-to-use forms, a browseable hierarchy, a thesaurusized index—to make sure the user isn't swamped; and measure how well an ontology provides a "good" basis for extracting semantic information?