Managing Information Extraction

[Tutorial Outline]

AnHai Doan[1], Raghu Ramakrishnan[2], Shivakumar Vaithyanathan[3]
[1]University of Illinois, [2]University of Wisconsin, [3]IBM Research at Almaden
anhai@cs.uiuc.edu, raghu@cs.wisc.edu, shiv@almaden.ibm.com

1. INTRODUCTION

Many applications increasingly involve a large amount of *unstructured data*. Examples of such data include email, text, Web pages, newsgroup postings, news articles, call-center text records, business reports, research papers, and so on. In its raw form, the data has limited value since we can do little with it beyond keyword search. Consequently, over the past two decades, significant efforts have focused on the problem of extracting structured information (e.g., researchers, publications, co-author and advising relationships, etc.) from such data. The extracted information is then exploited in search, browsing, querying, and mining.

In recent years, the explosion of unstructured data on the World-Wide Web has generated significant further interests in the above extraction problem, and helped position it as a central research goal in the database, AI, data mining, IR, NLP, and Web communities. An illustrative (but far from exhaustive) list of current projects that address this research goal include: (1) entity matching and approximate joins at AT&T Research, MSR and Stanford, (2) answering structured queries over text at Columbia and UCLA, (3) intelligent email and personal information management (PIM) at CMU, Massachusetts, MIT and Washington, (4) extracting and querying semantic entities/relations at IIT Bombay, CMU, MSR and Washington, (5) data cleaning at MSR, (6) doing OLAP-style analysis using extracted information at IBM Almaden and Wisconsin, (7) standardization efforts at IBM Watson on interfaces for NLP extraction tools, (8) managing unstructured data in bioinformatics at Illinois and Michigan, and (9) Web-based community information management (CIM) at Illinois and Wisconsin.

It is clear from the above list that the extraction problem has attracted wide interest in several research communities, and has been driven by a variety of applications. The current research efforts however have largely focused on developing specialized "blackbox" algorithms to address different aspects of the extraction problem in specific application contexts. Consequently, when targeting a different application context, typically the entire process of managing the unstructured data and the data extracted from it must be addressed from scratch, in an extremely labor-intensive and error-prone process. Only recently has a consensus started to build on the need for *re-usable tools*, and for *a unified management of the entire extraction process*, including extraction, storage, indexing, querying, and maintenance of both the original raw data and the extracted information.

In particular, we believe that such *unified management of extraction* is a logical next step in database support for text, going beyond integration of inverted indexes and support for keyword search in RDBMSs. It is a unique opportunity for the database community to extend the footprint of database systems to the most rapidly growing type of data, i.e., various forms of text, in a way that exploits the acknowledged strengths of database systems (queries over structured data and robust data management), while incorporating and extending extraction techniques developed in AI, IR and NLP. Success here is crucial to the acceptance of database systems as a repository for text corpora, as seamless extraction management would provide a compelling argument for moving text into a DBMS.

This tutorial makes the case for developing a unified framework for management of information extraction (IE). We:

1. Survey research on information extraction in the database, AI, NLP, IR, and Web communities in recent years.
2. Discuss why this is the right time for the database community to actively participate and address the problem of managing information extraction (including in particular the challenges of maintaining and querying the extracted information).
3. Show how interested researchers can take the next step, by pointing to open problems, available datasets, applicable standards, and software tools.

We do not assume prior knowledge of text management, NLP, extraction techniques, or machine learning.

2. TUTORIAL OUTLINE

Part 1: Motivating Applications

We discuss IE management for three real-world applications:

Business Intelligence : Consider an auto manufacturer who tracks customer service reports from multiple dealers. Each service report includes both structured attributes (e.g., date, customer ID, make, dealer name, etc.), and textual attributes (e.g., a "comments" field that records additional

SIGMOD 2006, June 27–29, 2006, Chicago, Illinois, USA.

information about the service). The manufacturer may want to ask "aggregate" questions such as *"What is the likelihood of brake problems in New York for Chevy vehicles whose service reports contain the name 'Kevin Jackson'?"* Such questions require us to combine information stored in the structured attributes with *semantic information extracted from the textual attributes.*

Community Information Management (CIM) : There are many *communities* on the Web, each focusing on a specific set of topics. Examples include communities of database researchers, movie goers, organization intranets, and online technical support groups. A database researcher may want to track all citations of a particular paper, or want to know all interesting connections between two researchers (e.g., do they share the same advisor?). To serve such information needs, we are developing a software platform to manage community information. Given a community, we first identify a set of relevant online data sources. Next, we crawl the sources at regular intervals (e.g., daily), and extract relevant entities and relationships (e.g., researchers, papers, advising, giving talks, etc.). Finally, we leverage the extracted entities and relations to provide user services such as browsing, keyword search, and structured querying.

Semantic Search : AVATAR Semantic Search tackles the problem of precision-oriented retrieval. A keyword query is interpreted as a set of *precise queries* in the context of information extracted from text. Consider the scenario of searching email. Suppose a user submits the keyword query [tom phone]. A standard keyword search engine will interpret this query as "*retrieve emails that contain the words tom and phone.*" However, such an interpretation will not return emails which mention a phone number but not the keyword "phone." AVATAR handles this, and more powerful semantic interpretations such as *retrieve emails sent by tom that mention a phone number*, by engaging the user in a dialog.

Part 2: State of the Art

1. *What are the steps in the IE management process?*

- *Structured data extraction:* Our survey includes (but is not limited to) rule- and learning- based information extraction approaches, recent efforts in the Statistical NLP community on name-entity recognition and simple relationship extraction, extraction efforts at Web scale, and extraction efforts in the database community.

- *Data cleaning & fusion:* Our focus is on topics relevant to managing information extraction. Examples include: (1) matching extracted entity mentions, such as 'J. N. Gray' and 'Jim Gray,' (2) merging inconsistent data, (3) verifying extracted information (e.g., that an extracted 'city name' is correct), and (4) merging outputs of multiple extraction systems.

- *Providing services over extracted data:* How can we execute keyword or SQL queries over extracted *structured* information? How can we mine extracted information?

- *Managing extracted data as the underlying raw data evolves:* We will discuss recent work on maintaining statistics over text databases, semantic matches over Deep-Web data sources, and mining models.

2. What problems and techniques play a fundamental role? We will survey a set of techniques, including managing uncertain data, data provenance, and handling data quality.

3. Where do things seem to be heading in the near term? We discuss current efforts to develop and improve "blackbox" algorithms to various problems in information extraction. Next, we discuss preliminary work in integrating certain "blackboxes," in particular, efforts on combining IE and entity matching and efforts on combining multiple IE systems. We also review attempts to standardize the API of black boxes, to ensure "plug and play," and discuss a growing awareness of additional issues that must be addressed within a unified framework: uncertainty management, provenance, scalability, exploiting user knowledge, and user interaction.

Part 3: Challenges & Opportunities

In this section we outline our perspectives on the major challenges and trends in this area. We discuss a next logical step: *developing a database management system to manage the entire process of information extraction.* We make the case for it, then discuss possible architectures and the associated challenges. If we are to design such a system, how should it look? What will be the capabilities? The key challenges that we discuss include data model and representational issues; need for newer index structures; standardization for IE; data cleaning and fusion; relationships between uncertainty management in the context of IE and probabilistic databases; data cleaning and fusion and finally the role of user knowledge and the iterative nature of user interaction. We ground the discussion in our current research efforts, but discuss numerous related and interlinked efforts across several research communities.

Part 4: How You Can Start

We believe that information extraction management provides many opportunities for a broad range of researchers in our community, for both the short and long term. It also provides an important unifying thread. Hence, in the final part of the tutorial, we discuss ways for database researchers to get started, with the lowest possible barrier of entry.

We first discuss multiple research themes, their challenges, and possible long-running competitions to build real-world applications that rely on IE management. We then survey data and real-world applications, as well as research materials (e.g., other tutorials, surveys, bibliographies) available to researchers.

About the presenters:

AnHai Doan has worked extensively on semantic integration and has co-edited special issues for SIGMOD Record 2004 and AI Magazine 2005 on semantic integration over structured data combined with text.

Raghu Ramakrishnan founded a company that developed collaborative customer support and investigated search over text and structured metadata. His research focuses on data retrieval, analysis, and mining.

Shivakumar Vaithyanathan leads the Unstructured Information Mining Group at the IBM Almaden Research Center. His primary research interest is in the area of machine learning algorithms, particularly unsupervised learning, with applications to text.