Managing Information Extraction
SIGMOD 2006 Tutorial

AnHai Doan
UIUC → UW-Madison

Raghu Ramakrishnan
UW-Madison → Yahoo! Research

Shiv Vaithyanathan
IBM Almaden
Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What’s different about managing IE?

- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback
The Presenters
AnHai Doan

- Currently at Illinois
- Starts at UW-Madison in July
- Has worked extensively in semantic integration, data integration, at the intersection of databases, Web, and AI
- Leads the Cimple project and builds DBLife in collaboration with Raghu Ramakrishnan and a terrific team of students
- Search for “anhai” on the Web
Raghu Ramakrishnan

- Research Fellow at Yahoo! Research, where he moved from UW-Madison after finding out that AnHai was moving there
- Has worked on data mining and database systems, and is currently focused on Web data management and online communities
- Collaborates with AnHai and gang on the Cimple/DBlife project, and with Shiv on aspects of Avatar
- See www.cs.wisc.edu/~raghu
Shiv Vaithyanathan

- Shiv Vaithyanathan manages the Unstructured Information Mining group at IBM Almaden where he moved after stints in DEC and Altavista.
- Shiv leads the Avatar project at IBM and is considering moving out of California now that Raghu has moved in.
- See www.almaden.ibm.com/software/projects/avatar/
Introduction
Lots of Text, Many Applications!

- Free-text, semi-structured, streaming …
  - Web pages, email, news articles, call-center text records, business reports, annotations, spreadsheets, research papers, blogs, tags, instant messages (IM), …

- High-impact applications
  - Business intelligence, personal information management, Web communities, Web search and advertising, scientific data management, e-government, medical records management, …

- Growing rapidly
  - Your email inbox!
Exploiting Text ➔
Important Direction for Our Community

- Many other research communities are looking at how to exploit text
  - Most actively, Web, IR, AI, KDD
- Important direction for us as well!
  - We have lot to offer, and a lot to gain
- How is text exploited?
  
  Two main directions: IR and IE
Exploiting Text via IR (Information Retrieval)

- Keyword search over data containing text (relational, XML)
  - What should the query language be? Ranking criteria?
  - How do we evaluate queries?

- Integrating IR systems with DB systems
  - Architecture?
  - See SIGMOD-04 panel; Baeza-Yates / Consens tutorial [SIGIR 05]

Not the focus of our tutorial
For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Select Name
From PEOPLE
Where Organization = ‘Microsoft’

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>Founder</td>
<td>Free Soft..</td>
</tr>
</tbody>
</table>

Bill Gates
Bill Veghte

(from Cohen’s IE tutorial, 2003)
This Tutorial: Research at the Intersection of IE and DB Systems

● We can apply DB approaches to
  – Analyzing and using extracted information in the context of other related data, as well as
  – The process of extracting and maintaining structured data from text

● A “killer app” for database systems?
  – Lots of text, but until now, mostly outside DBMSs
  – Extracted information could make the difference!

Let’s use three concrete applications to illustrate what we can do with IE …
A Disclaimer

This tutorial touches upon a lot of areas, some with much prior work. Rather than attempt a comprehensive survey, we’ve tried to identify areas for further research by the DB community.

We’ve therefore drawn freely from our own experiences in creating specific examples and articulating problems.

We are creating an annotated bibliography site, and we hope you’ll join us in maintaining it at http://scratchpad.wikia.com/wiki/Dblife_bibs
Application 1: Enterprise Search

Avatar Semantic Search
@ IBM Almaden
(and Shiv Vaithyanathan)
(SIGMOD Demo, 2006)
Overview of Avatar Semantic Search

Incorporate higher-level semantics into information retrieval to ascertain user-intent.

Conventional Search

- Query 1: return emails FROM Beineke that contain his contact telephone number
- Query 2: return emails that contain Beineke’s signature
- Query 3: return emails FROM Beineke that contain a telephone number

More ............

Avatar Semantic Search engages the user in a simple dialogue to ascertain user need.

Interpreted as

Return emails that contain the keywords “Beineke” and phone

It will miss

True user intent can be any of …
Results of the Semantic Optimizer
Blog Search Application

Two Interpretations of “hard rock”
How Semantic Search Works

- Semantic Search is basically KIDO (Keywords In Documents Out) enhanced by text-analytics.
- During offline processing, information extraction algorithms are used to extract specific facts from the raw text.
- At runtime, a “semantic optimizer” disambiguates the keyword query in the context of the extracted information and selects the best interpretations to present to the user.
Partial Type-System for Email
Concept tagged matches

**barbara matches**
1. value [Person.name]
2. value [Signature.person.name]
3. value [FromPhone.person.name]
4. value [Author.name]
5. keyword

**phone matches**
- type [PhoneNumber]
- path [FromPhone.phone]
- path [Signature.phone]
- path [NamePhone.phone]
- keyword

In the Enron E-mail connection the keyword query “barbara phone” has a total of 78 interpretations

*documents that contain an Author with name matching ‘barbara’ and a type PhoneNumber*
Application 2: Community Information Management (CIM)

Fei Chen
Pedro DeRose
Yoonkyong Lee
Warren Shen

The DBLife System @ Illinois / Wisconsin
(and AnHai Doan, Raghu Ramakrishnan)
Best-Effort, Collaborative Data Integration for Web Communities

- There are many data-rich communities
  - Database researchers, movie fans, bioinformatics
  - Enterprise intranets, tech support groups

- Each community = many disparate data sources + many people

- By integrating relevant data, we can enable search, monitoring, and information discovery:
  - Any interesting connection between researchers X and Y?
  - Find all citations of this paper in the past one week on the Web
  - What is new in the past 24 hours in the database community?
  - Which faculty candidates are interviewing this year, where?
  - What are current hot topics? Who has moved where?
Cimple Project @ Illinois/Wisconsin

Import & personalize data
Modify data, provide feedback

Keyword search
SQL querying
Question answering
Browse
Mining
Alerts, tracking
News summary
Prototype DBLife

Integrate data of the DB research community

- 1164 data sources
  - Crawled daily, 11000+ pages = 160+ MB / day
Data Extraction

Selected Publications
- "Learning from the Web to Match Deep-Web Query Interfaces, W. Wu, A. Doan, C. Yu. ICDE-06. PPT slides.
- "A Web-Based Software Collection: From Scientific Databases to SQL. A. Doan. ICDE-05. Poster.

This is DBLife's annotated version. http://www.cs.wisc.edu/dbworld

Call for Papers

AAAI Fall Symposium
Semantic Web for Collaboration
October 12-15, 2006
Arlington, VA

Deadline: June 15, 2006

Recent advances in computing and networking technologies have enabled a new generation of applications that are distributed and collaborate across organizational and geographic boundaries. These applications are characterized by the ability to share, integrate, and coordinate information from heterogeneous sources. Emerging technologies, such as the Semantic Web, promise to help realize this vision by providing a common framework for information sharing and integration. The AAAI Fall Symposium on the Semantic Web for Collaboration will focus on the challenges and opportunities presented by this vision. The symposium will bring together researchers, practitioners, and domain experts to discuss recent advances in Semantic Web technologies and their applications in collaboration scenarios.

Topics of interest include:

- Cyber-infrastructure
- Data integration
- Data warehousing
- Ontologies
- Peer-to-peer systems
- Semantic Web
- Web services
- XML
- XML-based data models

- Data Extraction
- Expert systems
- Feature extraction
- Machine learning
- Natural language processing
- Pattern recognition
- Semantic technologies
- Text mining
- Web-based systems

Call for Papers

The symposium will feature invited talks, technical papers, and posters. Technical papers should be 5-7 pages long and will be published in the AAAI symposium proceedings. Poster submissions should be 2-4 pages long and will be presented at the symposium.

Submissions should be sent as PDF files to: symp-submissions@aaai.org by June 15, 2006.

Acknowledgments

This work was partially supported by ONR grants N00014-05-1-0059 and N00014-06-1-0772.
Data Cleaning, Matching, Fusion

Raghu Ramakrishnan

co-authors = A. Doan, Divesh Srivastava, ...
Database Bibliographies by Topic

- active databases, constraint management
- applications and middleware
- approximation and uncertainty
- architecture, engines, and internals
- change management, maintenance
- data cleaning, data translation, data exchange, schema matching, record linkage
- data integration, heterogeneous database systems, interoperability
- data mining, classification, clustering
- data models, query languages, design analysis
- data reduction, compression, sampling
- data replication
- data storage, indexing, and access methods
- data warehousing and olap, decision support
- deductive databases, datalog
- derived data and materialized views
- extensibility and database evolution
This is DBLife's annotated version of

Georg Gottlob, Maurizio Lenzerini, Leonid Libkin, Rajeev Motwani, Marc

<table>
<thead>
<tr>
<th>Annotation 1</th>
<th>Proper name list (separator: ,)</th>
<th>Are these annotations correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Annotation 2</td>
<td>People mention (Go to superhomepage)</td>
<td>Yes</td>
</tr>
<tr>
<td>Annotation 3</td>
<td>People mention (Go to superhomepage)</td>
<td>Yes</td>
</tr>
<tr>
<td>Annotation 4</td>
<td>People mention (Go to superhomepage)</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Mass Collaboration

If enough users vote “not Divesh” on this picture, it is removed.
Current State of the Art

- Numerous domain-specific, hand-crafted solutions
  - imdb.com for movie domain
  - citeseer.com, dblp, rexa, Google scholar etc. for publication
  - techspec for engineering domain

- Very difficult to build and maintain, very hard to port solutions across domains

- The CIM Platform Challenge:
  - Develop a software platform that can be rapidly deployed and customized to manage data-rich Web communities
    - Creating an integrated, sustainable online community for, say, Chemical Engineering, or Finance, should be much easier, and should focus on leveraging domain knowledge, rather than on engineering details
Application 3: Scientific Data Management

AliBaba
@ Humboldt Univ. of Berlin
Summarizing PubMed Search Results

- PubMed/Medline
  - Database of paper abstracts in bioinformatics
  - 16 million abstracts, grows by 400K per year

- AliBaba: Summarizes results of keyword queries
  - User issues keyword query Q
  - AliBaba takes top 100 (say) abstracts returned by PubMed/Medline
  - Performs online entity and relationship extraction from abstracts
  - Shows ER graph to user

- For more detail
  - Contact Ulf Leser
  - System is online at http://wbi.informatik.hu-berlin.de:8080/
„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex.‘‘

\[ \text{CBF-A} \xleftarrow{\text{interact}} \text{complex} \rightarrow \text{CBF-C} \]

\[ \text{CBF-B} \xrightarrow{\text{associates}} \text{CBF-A-CBF-C complex} \]
**Z-100** is an *arabinomannan* extracted from *Mycobacterium tuberculosis* that has various immunomodulatory activities, such as the induction of *interleukin 12*, *interferon gamma* (*IFN-gamma*) and beta-chemokines. The effects of **Z-100** on *human immunodeficiency virus type 1* (*HIV-1*) replication in *human monocyte-derived macrophages* (*MDMs*) are investigated in this paper. In *MDMs*, **Z-100** markedly suppressed the replication of not only macrophage-tropic (M-tropic) *HIV-1* strain (*HIV-1JR-CSF*), but also *HIV-1* pseudotypes that possessed amphotropic *Moloney murine leukemia virus* or *vesicular stomatitis virus G* envelopes. **Z-100** was found to inhibit *HIV-1* expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly luciferase* gene) when this vector was transfected directly into *MDMs*. These findings suggest that **Z-100** inhibits virus replication, mainly at *HIV-1* transcription. However, **Z-100** also downregulated expression of the *cell surface* receptors CD4 and CCR5 in *MDMs* suggesting some inhibitory effect on *HIV-1* entry. Further experiments revealed that **Z-100** induced *IFN-beta* production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (*C/EBP*) beta transcription factor that represses *HIV-1* long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of *p38 mitogen-activated protein kinases* (*MAPK*), indicating that the *p38 MAPK* signalling pathway was involved in **Z-100**-induced repression of *HIV-1* replication in *MDMs*. These findings suggest that **Z-100** might be a useful immunomodulator for control of *HIV-1* infection.
The image contains a visual representation of PubMed through a query for the IFN gamma signaling pathway. The visual diagram includes nodes and edges representing various proteins and pathways such as IFN-alpha, IFN-beta, hypotalamus, and IFN-gamma. The extracted information includes text snippets from PubMed articles regarding IFN-gamma production and its role in normal cytotoxicity and the compartmentalization of IFN receptors in plasma membranes. Links to PubMed databases are also available for further exploration.
Feedback mode for community-curation
So we can do interesting and useful things with IE. And indeed there are many current IE efforts, and many with DB researchers involved

Still, these efforts have been carried out largely in isolation. In general, what does it take to build such an IE-based application?

Can we build a “System R” for IE-based applications?
To build a “System R” for IE applications, it turns out that

(1) It takes far more than what classical IE technologies offer
(2) Thus raising many open and important problems
(3) Several of which the DB community can address

The tutorial is about these three points
Tutorial Roadmap

• Introduction to managing IE [RR]
  – Motivation
  – What’s different about managing IE?

• Major research directions
  – Extracting mentions of entities and relationships [SV]
    – Uncertainty management
  – Disambiguating extracted mentions [AD]
    – Tracking mentions and entities over time
  – Understanding, correcting, and maintaining extracted data [AD]
    – Provenance and explanations
    – Incorporating user feedback
Managing Information Extraction

Challenges in Real-Life IE, and Some Problems that the Database Community Can Address
Let’s Recap Classical IE

- **Entity and relationship (link) extraction**
  - Typically, these are done at the document level
- **Entity resolution/matching**
  - Done at the collection-level
- **Efforts have focused mostly on**
  - Improving the accuracy of IE algorithms for extracting entities/links
  - Scaling up IE algorithms to large corpora

Real-world IE applications need more!

- **Complex IE tasks:** Although not the focus of this tutorial, there is much work on extracting more complex concepts
  - Events
  - Opinions
  - Sentiments
For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...
Classical IE: Entity Resolution (Mention Disambiguation / Matching)

- Common, because text is inherently ambiguous; must **disambiguate and merge** extracted data.
1) Complications in Extraction and Disambiguation

- Multi-step, user-guided workflows
  - In practice, developed iteratively
  - Each step must deal with uncertainty / errors of previous steps

- Integrating multiple data sources
  - Extractors and workflows tuned for one source may not work well for another source
  - Cannot tune extraction manually for a large number of data sources

- Incorporating background knowledge (e.g., dictionaries, properties of data sources, such as reliability/structure/patterns of change)

- Continuous extraction, i.e., monitoring
  - Challenges: Reconciling prior results, avoiding repeated work, tracking real-world changes by analyzing changes in extracted data
Complications in Understanding and Using Extracted Data

- Answering queries over extracted data, adjusting for extraction uncertainty and errors in a principled way
- Maintaining provenance of extracted data and generating understandable user-level explanations
- Incorporating user feedback to refine extraction/disambiguation
  - Want to correct specific mistake a user points out, and ensure that this is not “lost” in future passes of continuous monitoring scenarios
  - Want to generalize source of mistake and catch other similar errors (e.g., if Amer-Yahia pointed out error in extracted version of last name, and we recognize it is because of incorrect handling of hyphenation, we want to automatically apply the fix to all hyphenated last names)
Workflows in Extraction Phase

• Example: extract Person’s contact PhoneNumber

I will be out Thursday, but back on Friday. Sarah can be reached at 202-466-9160. Thanks for your help. Christi 37007.

Hand-coded: If a person-name is followed by “can be reached at”, then followed by a phone-number ➔ output a mention of the contact relationship

• A possible workflow
Workflows in Entity Resolution

- Workflows also arise in the matching phase
- As an example, we will consider two different matching strategies used to resolve entities extracted from collections of user home pages and from the DBLP citation website
  - The key idea in this example is that a more liberal matcher can be used in a simple setting (user home pages) and the extracted information can then guide a more conservative matcher in a more confusing setting (DBLP pages)
Example: Entity Resolution Workflow

\[ d_1: \text{Gravano’s Homepage} \]
L. Gravano, K. Ross.
Text Databases. SIGMOD 03
L. Gravano, J. Sanz.
Packet Routing. SPAA 91

\[ d_2: \text{Columbia DB Group Page} \]
Members
L. Gravano  K. Ross  J. Zhou
L. Gravano, J. Zhou.
Text Retrieval. VLDB 04

\[ d_3: \text{DBLP} \]
Luis Gravano, Kenneth Ross.
Digital Libraries. SIGMOD 04
Luis Gravano, Jingren Zhou.
Fuzzy Matching. VLDB 01
Luis Gravano, Jorge Sanz.
Packet Routing. SPAA 91

\[ d_4: \text{Chen Li’s Homepage} \]
C. Li.
Machine Learning. AAAI 04
C. Li, A. Tung.
Entity Matching. KDD 03

\[ s_0 \text{ matcher: Two mentions match if they share the same name.} \]

\[ s_1 \text{ matcher: Two mentions match if they share the same name and at least one co-author name.} \]
Since homepages are often unambiguous, we first match homepages using the simple matcher \( s_0 \). This allows us to collect co-authors for Luis Gravano and Chen Li.

So when we finally match with tuples in DBLP, which is more ambiguous, we (a) already have more evidence in the form (b) of co-authors, and (b) can use the more conservative matcher \( s_1 \).
Entity Resolution With Background Knowledge

... contact Ashish Gupta at UW-Madison ... (Ashish Gupta, UW-Madison)

Entity/Link DB

<table>
<thead>
<tr>
<th>A. K. Gupta</th>
<th><a href="mailto:agupta@cs.wisc.edu">agupta@cs.wisc.edu</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Koch</td>
<td><a href="mailto:koch@cs.uiuc.edu">koch@cs.uiuc.edu</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cs.wisc.edu</th>
<th>UW-Madison</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs.uiuc.edu</td>
<td>U. of Illinois</td>
</tr>
</tbody>
</table>

• Database of previously resolved entities/links
• Some other kinds of background knowledge:
  – “Trusted” sources (e.g., DBLP, DBworld) with known characteristics (e.g., format, update frequency)
Continuous Entity Resolution

- What if Entity/Link database is continuously updated to reflect changes in the real world? (E.g., Web crawls of user home pages)

- Can use the fact that few pages are new (or have changed) between updates. Challenges:
  - How much belief in existing entities and links?
  - Efficient organization and indexing
    - Where there is no meaningful change, recognize this and minimize repeated work
Continuous ER and Event Detection

The real world might have changed!
- And we need to detect this by analyzing changes in extracted information
Real-life IE: What Makes Extracted Information Hard to Use/Understand

- The extraction process is riddled with errors
  - How should these errors be represented?
  - Individual annotators are black-boxes with an internal probability model and typically output only the probabilities. While composing annotators how should their combined uncertainty be modeled?

- Semantics for queries over extracted data must handle the inherent ambiguity

- Lots of work
  - Classics: Fuhr-Rollecke; Imielinski-Lipski; ProbView; Halpern; …
  - Recent: See March 2006 Data Engineering bulletin for special issue on probabilistic data management (includes Green-Tannen survey/discussion of several proposals)
Some Recent Work on Uncertainty

- Many representations proposed, e.g.,
  - Confidence scores; Or-sets; Hierarchical imprecision

- Lots of recent work on querying uncertain data
  - E.g., Dalvi-Suciu identified classes of easy (PTIME) and hard (P#) queries and gave PTIME processing algorithms for easy ones
  - E.g., Burdick et al. (VLDB 05) considered single-table aggregations and showed how to assign confidence scores to hierarchically imprecise data in an intuitive way
  - E.g., Trio project (ICDE 06) considering how lineage can constrain the values taken by an imprecisely known object
  - E.g., Deshpande et al. (VLDB 04) consider data acquisition
  - E.g., Fagin et al. (ICDT 03) consider data exchange
Real-life IE: What Makes Extracted Information Hard to Use/Understand

- Users want to “drill down” on extracted data
  - We need to be able to explain the basis for an extracted piece of information when users “drill down”.
  - Many proof-tree based explanation systems built in deductive DB / LP /AI communities (Coral, LDL, EKS-V1, XSB, McGuinness, …)
  - Studied in context of provenance of integrated data (Buneman et al.; Stanford warehouse lineage, and more recently Trio)

- Concisely explaining complex extractions (e.g., using statistical models, workflows, and reflecting uncertainty) is hard
  - And especially useful because users are likely to drill down when they are surprised or confused by extracted data (e.g., due to errors, uncertainty).
System extracted “Gupta, D” using these rules:

(R1) David Gupta is a person name
(R2) If “first-name last-name” is a person name, then “last-name, f” is also a person name.

Knowing this, system builder can potentially improve extraction accuracy.

One way to do that:
(S1) Detect a list of items
(S2) If A straddles two items in a list ➔ A is not a person name

Incorrect. But why?
Real-life IE: What Makes Extracted Information Hard to Use/Understand

- Provenance becomes even more important if we want to leverage user feedback to improve the quality of extraction over time.
  - Maintaining an extracted “view” on a collection of documents over time is very costly; getting feedback from users can help
  - In fact, distributing the maintenance task across a large group of users may be the best approach
    - E.g., CIM
Incorporating Feedback

A. Gupta, D. Smith, Text mining, SIGMOD-06

System extracted “Gupta, D” as a person name

System extracted “Gupta, D” using rules:

(R1) David Gupta is a person name
(R2) If “first-name last-name” is a person name, then “last-name, f” is also a person name.

Knowing this, system can potentially improve extraction accuracy.

(1) Discover corrective rules such as S1—S2
(2) Find and fix other incorrect applications of R1 and R2

A general framework for incorporating feedback?
IE-Management Systems?

- In fact, everything about IE in practice is hard.

- Can we build a “System R for IE-in-practice”? *That’s* the grand challenge of “Managing IE”
  - **Key point:** Such a platform must provide support for the range of tasks we’ve described, yet be readily customizable to new domains and applications
System Challenges

- Customizability to new applications
- Scalability
- Detecting broken extractors
- Efficient handling of previously extracted information when components (e.g., annotators, matchers) are upgraded
- …
Customizable Extraction

- Cannot afford to implement extraction, and extraction management, from scratch for each application.
- What tasks can we abstract into a platform that can be customized for different applications?

What needs to be customizable?
- “Schema” level definition of entity and link concepts
- Extraction libraries
- Choices in how to handle uncertainty
- Choices in how to provide / incorporate feedback
- Choices in entity resolution and integration decisions
- Choices in frequency of updates, etc.
Scaling Up: Size is Just One Dimension!

- Corpus size
- Number of corpora
- Rate of change
- Size of extraction library
- Complexity of concepts to extract
- Complexity of background knowledge
- Complexity of guaranteeing uncertainty semantics when querying or updating extracted data
OK. But Why Now is the Right Time?
1. Emerging Attempts to Go Beyond Improving Accuracy of Single IE Algorithm

- Researchers are starting to examine
  - How to make blackboxes run efficiently [Sarawagi et al.]
  - How to integrate blackboxes
    - Combine IE and entity matching [McCallum etc.]
    - Combine multiple IE systems [Alpa et al.]

- Attempts to standardize API of blackboxes, to ensure plug and play
  - GATE, UIMA, etc.

- Growing awareness of previously mentioned issues
  - Uncertainty management / provenance
  - Scalability
  - Exploiting user knowledge / user interaction
  - Exploit extracted data effectively
2. Multiple Efforts to Build IE Applications, in Industry and Academia

- However, each in isolation
  - Citeseer, Cora, Rexa, Dblife, what else?
  - Numerous systems in industry
    - Web search engines use IE to add some semantics to search (e.g., recognize place names), and to do better ad placement
    - Enterprise search, business intelligence

- We should share knowledge now
Summary

- Lots of text, and growing …
- IE can help us to better leverage text
- Managing the entire IE process is important
- Lot of opportunities for the DB community
Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What’s different about managing IE?

- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback
Extracting Mentions of Entities and Relationships
Popular IE Tasks

● Named-entity extraction
  – Identify named-entities such as Persons, Organizations etc.

● Relationship extraction
  – Identify relationships between individual entities, e.g., Citizen-of, Employed-by etc.
  – e.g., Yahoo! acquired startup Flickr

● Event detection
  – Identifying incident occurrences between potentially multiple entities such Company-mergers, transfer-ownership, meetings, conferences, seminars etc.
But IE is Much, Much More ..

- Lesser known entities
  - Identifying rock-n-roll bands, restaurants, fashion designers, directions, passwords etc.

- Opinion / review extraction
  - Detect and extract informal reviews of bands, restaurants etc. from weblogs
  - Determine whether the opinions can be positive or negative
From: Shively, Hunter S.

Date: Tue, 26 Jun 2001 13:45:01 -0700 (PDT)

I-10W to exit 730 Peachridge RD (1 exit past Brookshire). Turn left on Peachridge RD. 2 miles down on the right--turquoise 'horses for sale' sign
I went to see the OTIS concert last night. 'Twas SO MUCH FUN I really had a blast. There were a bunch of other bands. I loved STAB (Sexually Transmitted Alcoholic Bastards). They were a really weird ska band and people were running around and...

I also went with Anya, Joven, and Morgan to see the OTIS concert last night. 'Twas SO MUCH FUN. I really had a blast. Sadly OTIS had very little to do with it. There were a bunch of other bands there playing and two in particular were amazing. I loved STAB (Sexually Transmitted Alcoholic Bastards), they were this really weird ska band and people were running around and skanking and jumping on each other. Joven and I skanked with them and got pushed around, that part was actually pretty fun even though some people at school will never look at me the same again. Lol. The sax player in that band was also hot, but that's a side note. They played their own versions of “I Will Survive” and the theme song from The Munsters (I love that show!). The other good band was my favorite and it's called Illusion (stupid name, but good!) and it's more like a hard rock band. We were all jumping around and freakin' out. It was great... oh yeah, and the lead singer... SOOOO hot... want to touch za hiney? Hehe, yeah... mostly everyone was too busy admiring him to think about where they were jumping.

...oh yeah, and OTIS was good too. Hehe... errm... yeah. I realized the biggest downside to going to see their shows is the people. I hate mostly everyone at the shows. It's like going to school where all the people I like have been vacuumsed out, except of course for the people I bring. I love you guys. Haha.

This morning: woke up and decided not to go the gym... too freakin' tired. Drove to Jamba Juice and tried wheatgrass juice for the first time. That shit is NASTY. Nobody try it! It's disgusting, it really does taste like grass. I figured there had to be a reason why people do wheatgrass shots. I thought it was all in movies. The place is still in business...
Intranet Web: Identify form-entry pages [Li et al, SIGIR, 2006]

1. To apply for Spring/Summer aid, a separate financial aid application called a **Spring/Summer Request for Funds (RFF)** is required. To apply for **Spring/Summer 2005** financial aid, contact the Office of Financial Aid.

The 2006 Spring/Summer Request for Funds (RFF) will be available on this website in late January 2006. Students must register for Spring/Summer classes in order to receive a financial aid offer.

2. You must also complete and submit to the federal processor a **Free Application for Federal Student Aid (FAFSA)** to apply for Spring/Summer aid, if you have not already done so.

> **For example:** If you are applying for 2006 Spring/Summer aid and you already submitted a 2005-2006 FAFSA to apply for Fall/Winter aid, do not submit another one to apply for 2006 Spring/Summer. However, if you did not submit a 2005-2006 FAFSA, you will need to submit one.
Intranet Web: Software download pages along with Software Name [Li et al, SIGIR, 2006]

Run the Sample Simulation: To ensure that the RapidPlayer plug-in is working correctly on your Windows computer, try this sample simulation after the plug-in is installed.

Macintosh OS X Users Need Citrix to Run Simulations

Macintosh OS X users do not need to install the RapidPlayer plug-in to use the M-Pathways training simulations. When Macintosh users enter the MAIS LINC URL, the system automatically launches an Internet Explorer browser that connects them directly to MAIS LINC. This browser is implemented through the Citrix ICA Client v. 6.30.323 for Macintosh OS X.

Verifying Unit Policies Regarding Software Installation

Many units, including LSA, Business and Finance, and the Hospital and Health Centers, do not want individuals to download software to their workstations from the Web. If you have questions about your unit’s policies regarding downloading software from the Web, contact your Unit Liaison.
Workflows in Extraction

I will be out Thursday, but back on Friday. **Sarah can be reached at 202-466-9160.**
Thanks for your help. Christi 37007.

Sarah’s phone is 202-466-9160

**Single-shot extraction**

Multi-step Workflow

Sara’s phone

Sarah

can be reached at

202-466-9160
Broadly-speaking two types of IE systems: hand-coded and learning-based.

What do they look like?
When best to use what?
Where can I learn more?

Lets start with hand-coded systems ...
Generic Template for hand-coded annotators

**Procedure Annotator** \((d, A_d)\)

- \(R_f\) is a set of rules to generate features
- \(R_g\) is a set of rules to create candidate annotations
- \(R_c\) is a set of rules to consolidate annotations created by \(R_g\)

1. \(Features = Compute\_Features(R_f, d)\)
2. \(\text{foreach } r \in R_g\)
   - \(Candidates = Candidates U ApplyRule(r, Features, A_d)\)
3. \(Results = Consolidate(R_c, Candidates)\)
   - return Results
Simplified Real Example in DBLife

- **Goal:** build a simple person-name extractor
  - input: a set of Web pages $W$, DB Research People Dictionary DBN
  - output: all mentions of names in DBN

- **Simplified DBLife Person-Name extraction**
  - **Obtain Features:** HTML tags, detect lists of proper-names
  - **Candidate Generation:**
    - for each name e.g., David Smith
      - generate variants (V): “David Smith”, “D. Smith”, “Smith, D.”, etc.
      - obtain candidate person-names in $W$ using V
  - **Consolidation:** if an occurrence straddles two proper-names then drop it
Compiled Dictionary

Candidate Generation Rule: Identifies Miller, R as a potential person’s name

Consolidation Rule: If a candidate straddles two elements of the list then drop it

Rule 1  This rule will find person names with a salutation (e.g. Dr. Laura Haas) and two capitalized words

<token> INITIAL </token>
<token> DOT </token>
<token> CAPSWORD </token>
<token> CAPSWORD </token>

Rule 2  This rule will find person names where two capitalized words are present in a Person dictionary

<token> PERSONDICT, CAPSWORD </token>
<token> PERSONDICT, CAPSWORD </token>

CAPSWORD : Word starting with uppercase, second letter lowercase
E.g., DeWitt will satisfy it (DEWITT will not)
\p{Upper}\p{Lower}[\p{Alpha}]{1,25}

DOT : The character ‘.’

Note that some names will be identified by both rules
Hand-coded rules can be arbitrarily complex

Find conference name in raw text

# Regular expressions to construct the pattern to extract conference names

# These are subordinate patterns
my $wordOrdinals="(?:first|second|third|fourth|fifth|sixth|seventh|eighth|ninth|tenth|eleventh|twelfth|thirteenth|fourteenth|fifteenth)";
my $numberOrdinals="(?:\d(?:1st|2nd|3rd|1th|2th|3th|4th|5th|6th|7th|8th|9th|0th))";
my $ordinals="(?:$wordOrdinals|$numberOrdinals)";
my $confTypes="(?:Conference|Workshop|Symposium)";
my $words="(?:[A-Z]w+.s*)"; # A word starting with a capital letter and ending with 0 or more spaces
my $confDescriptors="(?::international\s+[A-Z]+\s+)"; # e.g "International Conference ...' or the conference name for workshops (e.g. "VLDB Workshop ...")
my $connectors="(on|of)";
my $abbreviations="(?:\[(A-Z)w+\s*(?:\d\d+)?\]\s+)"; # Conference abbreviations like "(SIGMOD'06)"

# The actual pattern we search for. A typical conference name this pattern will find is # "3rd International Conference on Blah Blah Blah (ICBBB-05)"
my $fullNamePattern="((?:$ordinals\s+$words|$confDescriptors)\s*$confTypes(\s*$connectors\s+.?|$))\s*$abbreviations?\s*($)?";

# Given a <dbworldMessage>, look for the conference pattern
# lookForPattern($dbworldMessage, $fullNamePattern);

# In a given <file>, look for occurrences of <pattern>
# <pattern> is a regular expression
# sub lookForPattern { my ($file,$pattern) = @_;
Example Code of Hand-Coded Extractor

```perl
# Only look for conference names in the top 20 lines of the file
my $maxLines=20;
my $topOfFile=getTopOfFile($file,$maxLines);

# Look for the match in the top 20 lines - case insensitive, allow matches spanning multiple lines
if($topOfFile=~/(.*)$pattern/is) {
    my ($prefix,$name) = ($1,$2);
    # If it matches, do a sanity check and clean up the match
    # Get the first letter
    # Verify that the first letter is a capital letter or number
    if(!($name=~/^[A-Z0-9]/)) { return (); }
    # If there is an abbreviation, cut off whatever comes after that
    if($name=~/^(.*?$abbreviations)/) { $name=$1; }
    # If the name is too long, it probably isn't a conference
    if(scalar($name=~/\[^s\]/g) > 100) { return (); }
    # Get the first letter of the last word (need to this after chopping off parts of it due to abbreviation
    my ($letter,$nonLetter) = (^[A-Za-z]$letter*$nonLetter*$); # Need a space before $name to handle the first $nonLetter in the pattern if there
    # is only one word in name
    my $lastLetter=$1;
    if(!($lastLetter=~/^[A-Z]/)) { return (); } # Verify that the first letter of the last word is a capital letter

    # Passed test, return a new crutch
    return newCrutch(length($prefix),length($prefix)+length($name),$name,"Matched pattern in top $maxLines lines","conference name",getYear($name));
}
return ();
```

85
Some Examples of Hand-Coded Systems

- FRUMP [DeJong 82]
- CIRCUS / AutoSlog [Riloff 93]
- SRI FASTUS [Appelt, 1996]
- OSMX [Embley, 2005]
- DBLife [Doan et al, 2006]
- Avatar [Jayram et al, 2006]
Template for Learning based annotators

**Procedure LearningAnnotator (D, L)**

- D is the training data
- L is the labels

1. Preprocess D to extract features F
2. Use F,D & L to learn an extraction model E using a learning algorithm A
   *(Iteratively fine-tune parameters of the model and F)*

**Procedure ApplyAnnotator(d,E)**

1. Features = Compute.Features (d)
2. results = ApplyModel (E,Features, d)
3. return Results
Real Example in AliBaba

- Extract gene names from PubMed abstracts
- Use Classifier (Support Vector Machine - SVM)

- Corpus of 7500 sentences
  - 140,000 non-gene words
  - 60,000 gene names
- SVM\textsuperscript{light} on different feature sets
- Dictionary compiled from Genbank, HUGO, MGD, YDB
- Post-processing for compound gene names
Learning-Based Information Extraction

- Naive Bayes
- SRV [Freitag-98], Inductive Logic Programming
- Rapier [Califf & Mooney-97]
- Hidden Markov Models [Leek, 1997]
- Maximum Entropy Markov Models [McCallum et al, 2000]
- Conditional Random Fields [Lafferty et al, 2000]

For an excellent and comprehensive view [Cohen, 2004]
Semi-Supervised IE Systems
Learn to Gather More Training Data

1. Use labeled data to learn an extraction model E
2. Apply E to find mentions in document collection.
3. Construct more labeled data $\rightarrow T'$ is the new set.
4. Use $T'$ to learn a hopefully better extraction model $E'$.
5. Repeat.

[DIPE, Brin 98, Snowball, Agichtein & Gravano, 2000]
So there are basically two types of IE systems: hand-coded and learning-based.

What do they look like?
When best to use what?
Where can I learn more?
Hand-Coded Methods

- Easy to construct in many cases
  - e.g., to recognize prices, phone numbers, zip codes, conference names, etc.

- Easier to debug & maintain
  - especially if written in a “high-level” language (as is usually the case)
  - e.g.,

[From Avatar]

\[
\begin{align*}
\text{ContactPattern} & \leftarrow \text{RegularExpression(Email.body, ”can be reached at”)} \\
\text{PersonPhone} & \leftarrow \text{Precedes(Person, Precedes(Precedes(ContactPattern, Phone, D), D)}
\end{align*}
\]

- Easier to incorporate / reuse domain knowledge
- Can be quite labor intensive to write
Learning-Based Methods

- Can work well when training data is easy to construct and is plentiful
- Can capture complex patterns that are hard to encode with hand-crafted rules
  - e.g., determine whether a review is positive or negative
  - extract long complex gene names

[From AliBaba]

The human T cell leukemia lymphotrophic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300."*

- Can be labor intensive to construct training data
  - not sure how much training data is sufficient

Complementary to hand-coded methods
Where to Learn More

- **Overviews / tutorials**
  - Wendy Lehnert [Comm of the ACM, 1996]
  - Appelt [1997]
  - Cohen [2004]
  - Agichtein and Sarawai [KDD, 2006]
  - Andrew McCallum [ACM Queue, 2005]

- **Systems / codes to try**
  - OpenNLP
  - MinorThird
  - Weka
  - Rainbow
So what are the new IE challenges for IE-based applications?

First, let's discuss several observations, to motivate the new challenges.
Observation 1: We Often Need Complex Workflow

- What we have discussed so far are largely IE components
- Real-world IE applications often require a workflow that glue together these IE components
- These workflows can be quite large and complex
- Hard to get them right!
Illustrating Workflows

- Extract person’s contact phone-number from e-mail

I will be out Thursday, but back on Friday. Sarah can be reached at 202-466-9160. Thanks for your help. Christi 37007.

- A possible workflow

Hand-coded: If a person-name is followed by “can be reached at”, then followed by a phone-number ➔ output a mention of the contact relationship
How Workflows are Constructed

- Define the information extraction task
  - e.g., identify people’s phone numbers from email
- Identify the text-analysis components
  - E.g., tokenizer, part-of-speech tagger, Person, Phone annotator
- Compose different text-analytic components into a workflow
  - Several open-source plug-and-play architectures such as UIMA, GATE available
- Build domain-specific text-analytic component
How Workflows are Constructed

- Define the information extraction task
  - E.g., identify people’s phone numbers from email

- Identify the generic annotator components
  - E.g., tokenizer, part-of-speech tagger, Person, Phone annotator

- Compose different text-analytic components into a workflow
  - Several open-source plug-and-play architectures such as UIMA, GATE available

- Build domain-specific text-analytic component

Generic text-analytic tasks. Use available components
How Workflows are Constructed

● Define the information extraction task
  – E.g., identify people’s phone numbers from email

● Identify the text-analysis components
  – E.g., tokenizer, part-of-speech tagger, Person, Phone annotator

● Compose different text-analytic components into a workflow
  – Several open-source plug-and-play architectures such as UIMA, GATE available

● Build domain-specific text-analytic component
How Workflows are Constructed

- Define the information extraction task
  - E.g., identify people’s phone numbers from email

- Identify the generic text-analysis components
  - E.g., tokenizer, part-of-speech tagger, Person, Phone annotator

- Compose different text-analytic components into a workflow
  - Several open-source plug-and-play architectures such as UIMA, GATE available

- Build domain-specific text-analytic component
  - which is the contact relationship annotator in this example
Extracting Persons and Phone Numbers
Aggregate Analysis Engine: Person & Phone Detector

- Tokens
- Parts of Speech
- PhoneNumbers
- Persons

Tokenizer → Part of Speech … → Person And Phone Annotator → Relation Annotator

Identifying Person’s Phone Numbers from Email
Workflows are often Large and Complex

- In DBLife system
  - between 45 to 90 annotators
  - the workflow is 5 level deep
  - this makes up only half of the DBLife system (this is counting only extraction rules)

- In Avatar
  - 25 to 30 annotators extract a single fact with [SIGIR, 2006]
  - Workflows are 7 level deep
Observation 2: Often Need to Incorporate Domain Constraints

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular field.

Meeting: meeting(3:30pm, 5:00pm, Wean Hall)

start-time < end-time

if (location = “Wean Hall”) → start-time > 12

Meeting is from 3:30 – 5:00 pm in Wean Hall
Observation 3: The Process is Incremental & Iterative

- During development
  - Multiple versions of the same annotator might need to be compared and contrasted before choosing the right one (e.g., different regular expressions for the same task)
  - Incremental annotator development

- During deployment
  - Constant addition of new annotators; extract new entities, new relations etc.
  - Constant arrival of new documents
  - Many systems are 24/7 (e.g., DBLife)
Observation 4: Scalability is a Major Problem

- **DBLife example**
  - 120 MB of data / day, running the IE workflow once takes 3-5 hours
  - Even on smaller data sets debugging and testing is a time-consuming process
  - stored data over the past 2 years → magnifies scalability issues
  - write a new domain constraint, now should we rerun system from day one? Would take 3 months.

- **AliBaba: query time IE**
  - Users expect almost real-time response

Comprehensive tutorial - Sarawagi and Agichtein [KDD, 2006]
These observations lead to many difficult and important challenges
Efficient Construction of IE Workflow

- What would be the right workflow model?
  - Help write workflow quickly
  - Helps quickly debug, test, and reuse
  - UIMA / GATE? (do we need to extend these?)

- What is a good language to specify a single annotator in this workflow
  - An example of this is CPSL [Appelt, 1998]
  - What are the appropriate list of operators?
  - Do we need a new data-model?
  - Help users express domain constraints.
What are a good set of “operators” for IE process?
- Span operations e.g., Precedes, contains etc.
- Block operations
- Constraint handler?
- Regular expression and dictionary operators

Efficient implementation of these operators
- Inverted index constructor? inverted index lookup? [Ramakrishnan, G. et. al, 2006]

How to compile an efficient execution plan?
Finding a good execution plan is important!

Reuse existing annotations
- E.g., Person’s phone number annotator
- Lower-level operators can ignore documents that do NOT contain Persons and PhoneNumbers → potentially 10-fold speedup in Enron e-mail collection
- Useful in developing sparse annotators

Questions?
- How to estimate statistics for IE operators?
- In some cases different execution plans may have different extraction accuracy → not just a matter of optimizing for runtime
Rules as Declarative Queries in Avatar

Person can be reached at PhoneNumber

Person followed by ContactPattern followed by PhoneNumber

Declarative Query Language

ContactPattern ← RegularExpression(Email.body,"can be reached at")

PersonPhone ← Precedes (Precedes (Person, ContactPattern, D), Phone, D)
Domain-specific annotator in Avatar

- Identifying people’s phone numbers in email
  
  I will be out Thursday, but back on Friday. Sarah can be reached at 202-466-9160. Thanks for your help. Christi 37007.

- Generic pattern is
  
  Person can be reached at PhoneNumber
Optimizing IE Workflows in Avatar

- An IE workflow can be compiled into different execution plans
- E.g., two “execution plans” in Avatar:

Person can be reached at PhoneNumber

ContactPattern ← RegularExpression(Email.body,”can be reached at”)

PersonPhone ← Precedes (Precedes (Person, ContactPattern, D), Phone, D)

ContactPattern ← RegularExpression(Email.body,”can be reached at”)

PersonPhone ← Precedes(Person Precedes(ContactPattern, Phone, D), D)
### Alternative Query in Avatar

<table>
<thead>
<tr>
<th>ContactPattern</th>
<th>⇐</th>
<th>RegularExpression(Email.body,&quot;can be reached at&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PersonPhone</td>
<td>⇐</td>
<td>Contains ( Precedes (Person, Phone, D), ContactPattern )</td>
</tr>
</tbody>
</table>
I went to see the OTIS concert last night. T’ was SO MUCH FUN I really had a blast. There were a bunch of other bands. I loved STAB. They were a really weird ska band and people were running around and ...
Real challenge is in optimizing such complex workflows!!

Band INSTANCE PATTERNS

<Leading pattern> <Band instance> <Trailing pattern>

<MUSCIAN> <PERFORMED> <ADJECTIVE>
lead singer sang very well

<MUSCIAN> <ACTION> <INSTRUMENT>

Danny Sigelman played drums

<ADJECTIVE> <MUSIC>
energetic music

attended the Josh Groban concert at the Arrowhead

DESCRIPTION PATTERNS (Ambiguous/Unambiguous)

<Adjective> <Band or Associated concepts>

<Associated concept> <Linkage pattern> <Associated concept>

MUSIC, MUSICIANS, INSTRUMENTS, CROWD, ...
Yesterday, Went to see "The Piano," I finally thought it was good, it looked a lot for what it was, I didn't much care for the acting in the beginning, but towards the end they brought in some better actors and it was, well, better. I feel bad for the main actor as he seems to have gotten type cast as "Jewish" in every role he's played. I guess he must be the most "Jewish-looking" actor in Hollywood. Nice work if you can get it. I guess. The only exception was in Son of Sam where he played a transvestite... I'm not going to go there. Anyway, it was a good movie... it probably deserves Best Picture, it was really good. So far, "The Queen's" has been the one movie I'm going with as the best, whether or not they actually win. I need to post this before I end up reading some book review again... I hate being a movie nerd.

OTIS

Band instance pattern

(Un)ambiguous pattern

(Un)ambiguous pattern

(Un)ambiguous pattern

(Un)ambiguous pattern

(Un)ambiguous pattern

Review

Continuity
Tutorial Roadmap

- Introduction to managing IE [RR]
  - Motivation
  - What’s different about managing IE?

- Major research directions
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time
  - Understanding, correcting, and maintaining extracted data [AD]
    - Provenance and explanations
    - Incorporating user feedback
Uncertainty Management
Uncertainty During Extraction Process

- Annotators make mistakes!
- Annotators provide confidence scores with each annotation
- Simple named-entity annotator

\[ C = \text{Word with first letter capitalized} \]
\[ D = \text{Matches an entry in a person name dictionary} \]

Annotator Rules | Precision
--- | ---
1. [CD] [CD] | 0.9
2. [CD] | 0.6

Last evening I met the candidate Shiv Vaithyanathan for dinner. We had an interesting conversation and I encourage you to get an update. His host Bill can be reached at X-2465.

<table>
<thead>
<tr>
<th>Text-mention</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shiv Vaithyanathan</td>
<td>0.9</td>
</tr>
<tr>
<td>Bill</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Question: How do we compute probabilities for the output of composite annotators from base annotators?
With Two Annotators

Person Table

<table>
<thead>
<tr>
<th>ID</th>
<th>Text-mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shiv Vaithyanathan</td>
</tr>
<tr>
<td>2</td>
<td>Bill</td>
</tr>
</tbody>
</table>

Telephone Table

<table>
<thead>
<tr>
<th>ID</th>
<th>Text-mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(408)-927-2465</td>
</tr>
<tr>
<td>2</td>
<td>X-2465</td>
</tr>
</tbody>
</table>

These annotations are kept in separate tables
Last evening I met the candidate Shiv Vaithyanathan for dinner. We had an interesting conversation and I encourage you to get an update. His host Bill can be reached at X-2465.
One Potential Approach: Possible Worlds [Dalvi-Suciu, 2004]

Person example

<table>
<thead>
<tr>
<th>ID</th>
<th>Text-mention</th>
<th>0.9</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shiv Vaithyanathan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bill</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Text-Mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shiv Vaithyanathan</td>
</tr>
<tr>
<td>2</td>
<td>Bill</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Text-Mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shiv Vaithyanathan</td>
</tr>
<tr>
<td>2</td>
<td>Bill</td>
</tr>
</tbody>
</table>

ID 0.54

ID 0.36
Possible Worlds Interpretation [Dalvi-Suciu, 2004]

Bill appears in 60% of the possible worlds.

X-2465 appears in 30% of the possible worlds.

(Bill, X-2465) appears in at most 18% of the possible worlds.

Annotation (Bill, X-2465) can have a probability of at most 0.18.
With Enron collection using Person instances with a low probability the following rule

Person can be reached at PhoneNumber

produces annotations that are correct more than 80% of the time

Relaxing independence constraints [Fuhr-Roelleke, 95] does not help since X-2465 appears in only 30% of the worlds

More powerful probabilistic database constructs are needed to capture the dependencies present in the Rule above!
Databases and Probability

● Probabilistic DB
  – Fuhr [F&R97, F95]: uses events to describe possible worlds
  – [Dalvi&Suciu04]: query evaluation assuming independence of tuples
  – Trio System [Wid05, Das06]: distinguishes between data lineage and its probability

● Relational Learning
  – Bayesian Networks, Markov models: assumes tuples are independently and identically distributed
  – Probabilistic Relational Models [Koller+99]: accounts for correlations between tuples

● Uncertainty in Knowledge Bases
  – [GHK92, BGHK96] generating possible worlds probability distribution from statistics
  – [BGHK94] updating probability distribution based on new knowledge

● Recent work
Disambiguate, aka match, extracted mentions
Once mentions have been extracted, matching them is the next step.
Mention Matching: Problem Definition

- Given extracted mentions $M = \{m_1, \ldots, m_n\}$
- Partition $M$ into groups $M_1, \ldots, M_k$
  - All mentions in each group refer to the same real-world entity

- Variants are known as
  - Entity matching, record deduplication, record linkage, entity resolution, reference reconciliation, entity integration, fuzzy duplicate elimination
Another Example

**Document 1:** The Justice Department has officially ended its inquiry into the assassinations of John F. Kennedy and Martin Luther King Jr., finding "no persuasive evidence" to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that Kennedy was "probably" assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the Warren Commission's belief that Lee Harvey Oswald acted alone in Dallas on Nov. 22, 1963.

**Document 2:** In 1953, Massachusetts Sen. John F. Kennedy married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate John F. Kennedy confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, "I do not speak for my church on public matters, and the church does not speak for me."


[From Li, Morie, & Roth, AI Magazine, 2005]
Extremely Important Problem!

- Appears in numerous real-world contexts
- Plagues many applications that we have seen
  - Citeseer, DBLife, AliBaba, Rexa, etc.

Why so important?

- Many useful services rely on mention matching being right
- If we do not match mentions with sufficient accuracy
  - errors cascade, greatly reducing the usefulness of these services
An Example

Discover related organizations using occurrence analysis:

“J. Han ... Centrum voor Wiskunde en Informatica”

DBLife incorrectly matches this mention “J. Han” with “Jiawei Han”, but it actually refers to “Jianchao Han”.
The Rest of This Section

● To set the stage, briefly review current solutions to mention matching / record linkage
  – a comprehensive tutorial is provided tomorrow Wed 2-5:30pm, by Nick Koudas, Sunita Sarawagi, & Divesh Srivastava

● Then focus on novel challenges brought forth by IE over text
  – developing matching workflow, optimizing workflow, incorporating domain knowledge
  – tracking mentions / entities, detecting interesting events
A First Matching Solution: String Matching

\[ m_{11} = \text{“John F. Kennedy”} \]
\[ m_{12} = \text{“Kennedy”} \]

\[ m_{21} = \text{“Senator John F. Kennedy”} \]
\[ m_{22} = \text{“John F. Kennedy”} \]

\[ m_{31} = \text{“David Kennedy”} \]
\[ m_{32} = \text{“Kennedy”} \]

\[ \text{sim}(m_i, m_j) > 0.8 \Rightarrow m_i \text{ and } m_j \text{ match.} \]

\[ \text{sim} = \text{edit distance, q-gram, TF/IDF, etc.} \]

- A recent survey:
  - Other recent work: [Koudas, Marathe, Srivastava, VLDB-04]

- Pros & cons
  - conceptually simple, relatively fast
  - often insufficient for achieving high accuracy
A More Common Solution

- For each mention m, extract additional data
  - transform m into a record
- Match the records
  - leveraging the wealth of existing record matching solutions


<table>
<thead>
<tr>
<th>first-name</th>
<th>last-name</th>
<th>birth-date</th>
<th>birth-place</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>Kennedy</td>
<td>1959</td>
<td>Leicester</td>
</tr>
<tr>
<td>D.</td>
<td>Kennedy</td>
<td>1959</td>
<td>England</td>
</tr>
</tbody>
</table>
Two main groups of record matching solutions

- hand-crafted rules
- learning-based

which we will discuss next
### Hand-Crafted Rules

- If $R_1$.last-name = $R_2$.last-name
- $R_1$.first-name ~ $R_2$.first-name
- $R_1$.address ~ $R_2$.address

$\Rightarrow$ $R_1$ matches $R_2$

[Hernandez & Stolfo, SIGMOD-95]

\[
\text{sim}(R_1,R_2) = \alpha_1 \times \text{sim}_1(R_1\.last\text{-}name,R_2\.last\text{-}name) + \\
\alpha_2 \times \text{sim}_2(R_1\.first\text{-}name,R_2\.first\text{-}name) + \\
\alpha_3 \times \text{sim}_3(R_1\.address,R_2\.address)
\]

If \( \text{sim}(R_1,R_2) > 0.7 \) $\Rightarrow$ match

- **Pros and cons**
  - relatively easy to craft rules in many cases
  - easy to modify, incorporate domain knowledge
  - laborious tuning
  - in certain cases may be hard to create rules manually
Learning-Based Approaches

- Learn matching rules from training data
- Create a set of features: $f_1, \ldots, f_k$
  - each feature is a function over $(t,u)$
  - e.g., $t$.last-name = $u$.last-name?
  - edit-distance($t$.first-name, $u$.first-name)
- Convert each tuple pair to a feature vector, then apply a machine learning algorithm

\[
\begin{align*}
(t_1, u_1, +) \\
(t_2, u_2, +) \\
(t_3, u_3, -) \\
\vdots \\
(t_n, u_n, +)
\end{align*}
\]

\[
\begin{align*}
([f_{11}, \ldots, f_{1k}], +) \\
([f_{21}, \ldots, f_{2k}], +) \\
([f_{31}, \ldots, f_{3k}], -) \\
\vdots \\
([f_{n1}, \ldots, f_{nk}], +)
\end{align*}
\]

Decision tree, Naive Bayes, SVM, etc. \rightarrow \text{Learned “rules”}
Example of Learned Matching Rules

- Produced by a decision-tree learner, to match paper citations

[YearDifference > 1]

- All-Ngrams ≤ 0.48
  - Non-Duplicate
    - AuthorTitleNgrams ≤ 0.4
      - Duplicate
    - Duplicate
  - Duplicate
  - Duplicate
  - Non-Duplicate

[Sarawagi & Bhamidipaty, KDD-02]
Twists on the Basic Methods

- Compute transitive closures
  - [Hernandez & Stolfo, SIGMOD-95]

- Learn all sorts of other thing (not just matching rules)
  - e.g., transformation rules [Tejada, Knoblock, & Minton, KDD-02]

- Ask users to label selected tuple pairs (active learning)
  - [Sarawagi & Bhamidipaty, KDD-02]

- Can we leverage relational database?
  - [Gravano et. al., VLDB-01]
Twists on the Basic Methods

- Record matching in data warehouse contexts
  - Tuples can share values for subsets of attributes
  - [Ananthakrishna, Chaudhuri, & Ganti, VLDB-02]

- Combine mention extraction and matching
  - [Wellner et. al., UAI-04]

- And many more
  - e.g., [Jin, Li, Mehrotra, DASFAA-03]
  - TAILOR record linkage project at Purdue [Elfeky, Elmagarmid, Verykios]
Collective Mention Matching: A Recent Trend

- Prior solutions
  - assume tuples are immutable (can’t be changed)
  - often match tuples of just one type

- Observations
  - can enrich tuples along the way $\Rightarrow$ improve accuracy
  - often must match tuples of interrelated types $\Rightarrow$ can leverage matching one type to improve accuracy of matching other types

- This leads to a flurry of recent work on collective mention matching
  - which builds upon the previous three solution groups

- Will illustrate enriching tuples
  - Using [Li, Morie, & Roth, AAAI-04]
Example of Collective Mention Matching

1. Use a simple matching measure to cluster mentions in each document. Each cluster $\rightarrow$ an entity. Then learn a “profile” for each entity.

2. Reassign each mention to the best matching entity.

3. Recompute entity profiles. 4. Repeat Steps 2-3 until convergence.
Collective Mention Matching

1. Match tuples
2. “Enrich” each tuple with information from other tuples that match it; or create “super tuples” that represent groups of matching tuples.
3. Repeat Steps 1-2 until convergence.

Key ideas: enrich each tuple, iterate

Some recent algorithms that employ these ideas:

Pedro Domingos group at Washington, Dan Roth group at Illinois, Andrew McCallum group at UMass, Lise Getoor group at Maryland, Alon Halevy group at Washington (SEMEX), Ray Mooney group at Texas-Austin, Jiawei Han group at Illinois, and more
What new mention matching challenges does IE over text raise?

1. **Static data**: challenges similar to those in extracting mentions.

2. **Dynamic data**: challenges in tracking mentions / entities
Classical Mention Matching

- Applies just a single “matcher”
- Focuses mainly on developing matchers with higher accuracy

Real-world IE applications need more
To illustrate with a simple example:

<table>
<thead>
<tr>
<th>d₁: Luis Gravano’s Homepage</th>
<th>d₂: Columbia DB Group Page</th>
<th>d₃: DBLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. Gravano, K. Ross.</td>
<td>Members</td>
<td>Luis Gravano, Kenneth Ross.</td>
</tr>
<tr>
<td>Text Databases. SIGMOD 03</td>
<td>L. Gravano K. Ross J. Zhou</td>
<td>Digital Libraries. SIGMOD 04</td>
</tr>
<tr>
<td>Packet Routing. SPAA 91</td>
<td>Text Retrieval. VLDB 04</td>
<td>Fuzzy Matching. VLDB 01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Luis Gravano, Jorge Sanz.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Packet Routing. SPAA 91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chen Li, Anthony Tung.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Entity Matching. KDD 03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chen Li, Chris Brown.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interfaces. HCI 99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d₄: Chen Li’s Homepage</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Li.</td>
</tr>
<tr>
<td>Machine Learning. AAAI 04</td>
</tr>
<tr>
<td>C. Li, A. Tung.</td>
</tr>
<tr>
<td>Entity Matching. KDD 03</td>
</tr>
</tbody>
</table>

What is the best way to match mentions here?

Only one Luis Gravano

Two Chen Li-s
A liberal matcher: correctly predicts that there is one Luis Gravano, but incorrectly predicts that there is one Chen Li.

$s_0$ matcher: two mentions match if they share the same name.

- **d₁**: Luis Gravano’s Homepage
  - L. Gravano, K. Ross.
  - Text Databases. SIGMOD 03
  - L. Gravano, J. Sanz.
  - Packet Routing. SPAA 91

- **d₂**: Columbia DB Group Page
  - Members
  - L. Gravano, K. Ross, J. Zhou
  - L. Gravano, J. Zhou.
  - Text Retrieval. VLDB 04

- **d₃**: DBLP
  - Luis Gravano, Kenneth Ross.
  - Digital Libraries. SIGMOD 04
  - Luis Gravano, Jingren Zhou.
  - Fuzzy Matching. VLDB 01
  - Luis Gravano, Jorge Sanz.
  - Packet Routing. SPAA 91

- **d₄**: Chen Li’s Homepage
  - C. Li.
  - Machine Learning. AAAI 04
  - C. Li, A. Tung.
  - Entity Matching. KDD 03
  - Chen Li, Anthony Tung.
  - Entity Matching. KDD 03
  - Chen Li, Chris Brown.
  - Interfaces. HCI 99
A conservative matcher: predicts multiple Gravanos and Chen Lis

### s₁ matcher: two mentions match if they share the same name and at least one co-author name.

<table>
<thead>
<tr>
<th>d₁: Luis Gravano’s Homepage</th>
<th>d₂: Columbia DB Group Page</th>
<th>d₃: DBLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Luis Gravano, Jorge Sanz. Packet Routing. SPAA 91</td>
</tr>
<tr>
<td>d₄: Chen Li’s Homepage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Li. Machine Learning. AAAI 04</td>
<td></td>
<td>Chen Li, Anthony Tung. Entity Matching. KDD 03</td>
</tr>
<tr>
<td>C. Li, A. Tung. Entity Matching. KDD 03</td>
<td></td>
<td>Chen Li, Chris Brown. Interfaces. HCI 99</td>
</tr>
</tbody>
</table>
Better solution: apply both matchers in a workflow

\[ d_1: \text{Luis Gravano's Homepage} \]
L. Gravano, K. Ross.
Text Databases. SIGMOD 03
L. Gravano, J. Sanz.
Packet Routing. SPAA 91

\[ d_2: \text{Columbia DB Group Page} \]
Members
L. Gravano, K. Ross, J. Zhou
L. Gravano, J. Zhou.
Text Retrieval. VLDB 04

\[ d_3: \text{DBLP} \]
Luis Gravano, Kenneth Ross.
Digital Libraries. SIGMOD 04
Luis Gravano, Jingren Zhou.
Fuzzy Matching. VLDB 01
Luis Gravano, Jorge Sanz.
Packet Routing. SPAA 91

\[ d_4: \text{Chen Li's Homepage} \]
C. Li.
Machine Learning. AAAI 04
C. Li, A. Tung.
Entity Matching. KDD 03

\[ s_0 \]
union
[union]

\[ s_1 \]

\[ s_0 \] \quad \text{matcher: two mentions match if they share the same name.} \]

\[ s_0 \] \quad \text{matcher: two mentions match if they share the same name and at least one co-author name.}
We control how tuple enrichment happens, using different matchers.

Since homepages are often unambiguous, we first match homepages using the simple matcher $s_0$. This allows us to collect co-authors for Luis Gravano and Chen Li.

So when we finally match with tuples in DBLP, which is more ambiguous, we (a) already have more evidence in form of co-authors, and (b) use the more conservative matcher $s_1$. 
Another Example

- Suppose distinct researchers X and Y have very similar names, and share some co-authors
  - e.g., Ashish Gupta and Ashish K. Gupta
- Then $s_1$ matcher does not work, need a more conservative matcher $s_2$

```
union
  |
  s_1

union
  |
  d_3

union
  |
  s_0

union
  |
  d_4

union
  |
  s_0

All mentions with last name = Gupta
```
Need to Exploit a Lot of Domain Knowledge in the Workflow

[From Shen, Li, Doan, AAAI-05]

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate</td>
<td>No researcher has chaired more than 3 conferences in a year</td>
</tr>
<tr>
<td>Subsumption</td>
<td>If a citation X from DBLP matches a citation Y in a homepage, then each author in Y matches some author in X</td>
</tr>
<tr>
<td>Neighborhood</td>
<td>If authors X and Y share similar names and some coauthors, they are likely to match</td>
</tr>
<tr>
<td>Incompatible</td>
<td>No researcher exists who has published in both HCI and numerical analysis</td>
</tr>
<tr>
<td>Layout</td>
<td>If two mentions in the same document share similar names, they are likely to match</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>Mentions in the PC listing of a conference refer to different researchers</td>
</tr>
<tr>
<td>Ordering</td>
<td>If two citations match, then their authors will be matched in order</td>
</tr>
<tr>
<td>Individual</td>
<td>The researcher named “Mayssam Saria” has fewer than five mentions in DBLP (e.g. being a new graduate student with fewer than five papers)</td>
</tr>
</tbody>
</table>
Need Support for Incremental update of matching workflow

- We have run a matching workflow $E$ on a huge data set $D$
- Now we modified $E$ a little bit into $E'$
- How can we run $E'$ efficiently over $D$?
  - exploiting the results of running $E$ over $D$
- Similar to exploiting materialized views
- Crucial for many settings:
  - testing and debugging
  - expansion during deployment
  - recovering from crash
Research Challenges

● Similar to those in extracting mentions
● Need right model / representation language
● Develop basic operators: matcher, merger, etc.
● Ways to combine them \(\rightarrow\) match execution plan

● Ways to optimize plan for accuracy/runtime
  – challenge: estimate their performance
● Akin to relational query optimization
The Ideal Entity Matching Solution

- We throw in all types of information
  - training data (if available)
  - domain constraints
- and all types of matchers + other operators
  - SVM, decision tree, etc.
- Must be able to do this as declaratively as possible (similar to writing a SQL query)
- System automatically compile a good match execution plan
  - with respect to accuracy/runtime, or combination thereof
- Easy for us to debug, maintain, add domain knowledge, add patches
Recent Work / Starting Point

• SERF project at Stanford
  – Develops a generic infrastructure
  – Defines basic operators: match, merge, etc.
  – Finds fast execution plans

• Data cleaning project at MSR
  – Solution to match incoming records against existing groups
  – E.g., [Chaudhuri, Ganjam, Ganti, Motwani, SIGMOD-03]

• Cimple project at Illinois / Wisconsin
  – SOCCER matching approach
  – Defines basic operators, finds highly accurate execution plans
  – Methods to exploit domain constraints [Shen, Li, Doan, AAAI-05]

• Semex project at Washington
  – Methods to exploit domain constraints [Dong et. al., SIGMOD-05]
Mention Tracking

How do you tell if a mention is old or new?
- Compare mention semantics between days
- How do we determine a mention’s semantics?

John Smith’s Homepage

John Smith is a Professor at Foo University.
...

Selected Publications:
• ComPLEX. B. Santos, J. Smith.
• Databases and Me: C. Wu, D. Sato, J. Smith.
...

day n

John Smith’s Homepage

John Smith is a Professor at Bar University.
...

Selected Publications:
• Databases and That One Guy. J. Smith.
• ComPLEX: Not So Simple. B. Santos, J. Smith.
• Databases and Me. C. Wu, D. Sato, J. Smith.
...

day n+1
**Mention Tracking**

- Using fixed-width context windows often works …

- **But not always.**
  - ComPLEX. B. Santos, J. Smith.

- **Even intelligent windows can use help with semantics**
  - Databases and Me. C. Wu, D. Sato, J. Smith.
Entity Tracking

- Like mention tracking, how do you tell if an entity is old or new?
- Entities are sets of mentions, so we use a Jaccard distance:

\[
\frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}
\]

where \(E_1\) and \(E_2\) are entities on days \(k\) and \(k+1\) respectively.

<table>
<thead>
<tr>
<th>Day (k)</th>
<th>Day (k+1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entity E1</strong></td>
<td><strong>Entity F1</strong></td>
</tr>
<tr>
<td>m1, m2</td>
<td>n1, n2, n3</td>
</tr>
<tr>
<td><strong>Entity E2</strong></td>
<td><strong>Entity F2</strong></td>
</tr>
<tr>
<td>m3, m4, m5</td>
<td>m3, m4, m5</td>
</tr>
</tbody>
</table>
Monitoring and Event Detection

- The real world might have changed!
  - And we need to detect this by analyzing changes in extracted information

Infer that Raghu Ramakrishnan has moved to Yahoo! Research
Tutorial Roadmap

- **Introduction to managing IE** [RR]
  - Motivation
  - What’s different about *managing* IE?

- **Major research directions**
  - Extracting mentions of entities and relationships [SV]
    - Uncertainty management
  - Disambiguating extracted mentions [AD]
    - Tracking mentions and entities over time

Understanding, correcting, and maintaining extracted data [AD]
  - Provenance and explanations
  - Incorporating user feedback
Understanding, Correcting, and Maintaining Extracted Data
Understanding Extracted Data

- **Important in at least three contexts**
  - Development ➔ developers can fine tune system
  - Provide services (keyword search, SQL queries, etc.) ➔ users can be confident in answers
  - Provide feedback ➔ developers / users can provide good feedback

- **Typically provided as provenance (aka lineage)**
  - Often a tree showing the origin and derivation of data
An Example

System extracted contact(Sarah, 202-466-9160). Why?

I will be out Thursday, but back on Friday. Sarah can be reached at 202-466-9160. Thanks for your help. Christi 37007.

This rule fired:
person-name + “can be reached at” + phone-number ➔ output a mention of the contact relationship

Used regular expression to recognize “202-466-9160” as a phone number
In Practice, Need More than Just Provenance Tree

- Developer / user often want **explanations**
  - why X was extracted?
  - why Y was not extracted?
  - why system has higher confidence in X than in Y?
  - what if ... ?

- Explanations thus are related to, but different from provenance
An Example

I will be out Thursday, but back on Friday. Sarah can be reached at 202-466-9160. Thanks for your help. Christi 37007.

Why was “202-466-9160” not extracted?

Explanation:
(1) The relationship annotator uses the following rule to extract 37007:
person name + at most 10 tokens +
“can be reached at” +
at most 6 tokens + phone number \(\rightarrow\) contact(person name, phone number).

(2) “202-466-9160” fits into the part “at most 6 tokens”.
Generating Explanations is Difficult

- Especially for
  - why was A not extracted?
  - why does system rank A higher than B?

- Reasons
  - many possible causes for the fact that “A was not extracted”
  - must examine the provenance tree to know which components are chiefly responsible for causing A to be ranked higher than B
  - provenance trees can be huge, especially in continuously running systems, e.g., DBLife

- Some work exist in related areas, but little on generating explanations for IE over text
  - see [Dhamankar et. al., SIGMOD-04]:
    generating explanations for schema matching
System developers and users can use explanations / provenance to provide feedback to system (i.e., this extracted data piece is wrong), or manually correct data pieces.

This raises many serious challenges.

Consider the case of multiple users’ providing feedback ...
Motivating Example
The General Idea

- Many real-world applications inevitably have multiple developers and many users
- How to exploit feedback efforts from all of them?
- Variants of this is known as
  - collective development of system, mass collaboration, collective curation, Web 2.0 applications, etc.
- Has been applied to many applications
  - open-source software, bug detection, tech support group, Yahoo! Answers, Google Co-op, and many more
- Little has been done in IE contexts
  - except in industry, e.g., epinions.com
Challenges

- If X and Y both edit a piece of extracted data D, they may edit the same data unit differently.
- How would X and Y reconcile / share their edition?
- E.g., the ORCHESTRA project at Penn [Taylor & Ives, SIGMOD-06]

- How to entice people to contribute?
- How to handle malicious users?
- What types of extraction tasks are most amenable to mass collaboration?
- E.g., see MOBS project at Illinois [WebDB-03, ICDE-05]
As data evolves, extractors often break

<HTML>
<TITLE>Some Country Codes</TITLE>
<B>Congo</B>  <I>242</I>  <BR>
<B>Egypt</B>  <I>20</I>  <BR>
<B>Belize</B>  <I>501</I>  <BR>
<B>Spain</B>  <I>34</I>  <BR>
</BODY></HTML>

(Congo, 242)
(Egypt, 20)
(Belize, 501)
(Spain, 34)

<HTML>
<TITLE>Some Country Codes</TITLE>
<B>Congo</B>  <I>Africa</I>  <I>242</I>  <BR>
<B>Egypt</B>  <I>Africa</I>  <I>20</I>  <BR>
<B>Belize</B>  <I>N. America</I>  <I>501</I>  <BR>
<B>Spain</B>  <I>Europe</I>  <I>34</I>  <BR>
</BODY></HTML>

(Congo, Africa)
(Egypt, Africa)
(Belize, N. America)
(Spain, Europe)
Maintenance: Key Challenges

- Detect if an extractor or a set of extractors is broken
- Pinpoint the source of errors
- Suggest repairs or automatically repairs extractors
- Build semantic debuggers?
- Scalability issues
Related Work / Starting Points

- **Detect broken extractors**
  - Nick Kushmerick group in Ireland, Craig Knoblock group at ISI, Chen Li group at UCI, AnHai Doan group at Illinois

- **Repair broken extractors**
  - Craig Knoblock group at ISI

- **Mapping maintenance**
  - Renee Miller group at Toronto, Lucian Popa group at Almaden
Summary: Key Points of Tutorial

- Lot of future activity in text / Web management

- To build IE-based applications ➔ must go beyond developing IE components, to **managing the entire IE process**:
  - Manage the IE workflow, manage mention matching
  - Provide useful services over extracted data
  - Manage uncertainty, understand, correct, and maintain extracted data

- Solutions here + IR components ➔ can significantly extend the footprint of DBMSs

Think “System R” for IE-based applications!
How Can You Start

- We are putting pointers to literature, tools, & data at [http://scratchpad.wikia.com/wiki/Dblife_bibs](http://scratchpad.wikia.com/wiki/Dblife_bibs) (all current DBLife bibliographies also reside here)

- Please contribute!
- Also watch that space
  - Tutorial slides will be put there
  - Data will be available from DBLife, Avatar project, and Yahoo, in significant amount

- Will be able to navigate there from our homepages