A Scalable Architecture for Extracting, Aligning, Linking, and Visualizing Multi-Int Data

Craig Knoblock & Pedro Szekely

University of Southern California

Introduction

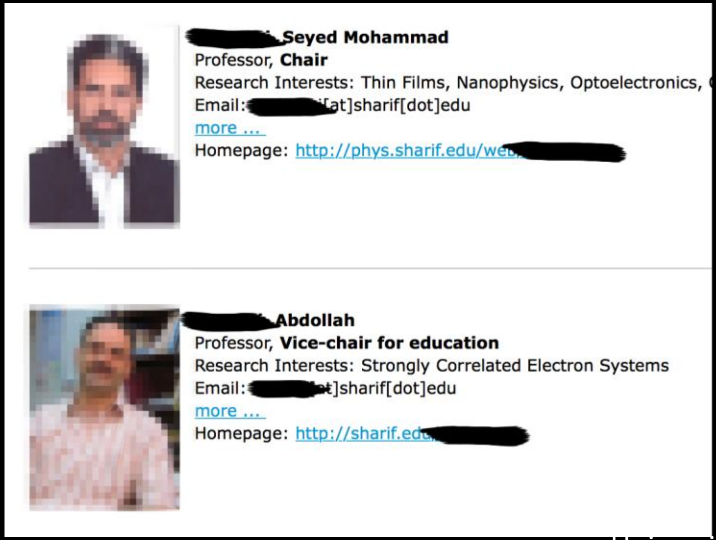
- Massive quantities of data available for analysis
  - OSING, HUMINT, SIGINT, MASINT, GEOINT, …
- Data is spread across multiple sources, multiple sites and multiple formats
  - Databases, text, web sites, XML, JSON, etc...
- If an analyst could exploit all of this data, it could transform analysis
  - Disruptive technology for analysis

Solution:
Domain-specific Insight Graphs

- Innovative architecture
  - Extracting, aligning, linking, and visualizing massive amounts of data
  - Domain-specific content from structured and unstructured sources
- State-of-the art open source software
  - Open architecture with flexible APIs
  - Cloud-based infrastructure (HDFS, Hadoop, ElasticSearch, etc.)

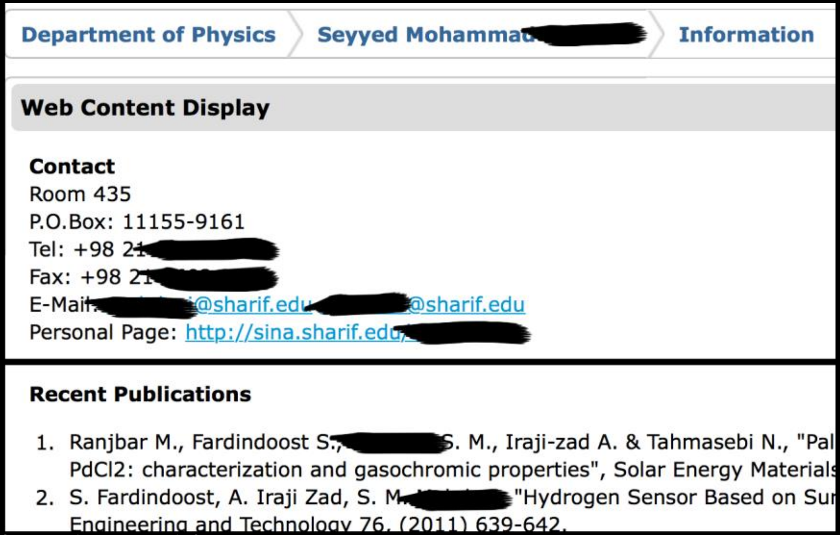
Example Scenario

- Want to determine the nuclear know-how of a given country from open source data
- Analyze the universities, academics, publications, reports, articles within the country



[redacted] Seyed Mohammad
Professor, **Chair**
Research Interests: Thin Films, Nanophysics, Optoelectronics,
Email:[redacted]at]sharif[dot]edu
more ...
Homepage: http://phys.sharif.edu/we[redacted]

[redacted] Abdollah
Professor, **Vice-chair for education**
Research Interests: Strongly Correlated Electron Systems
Email:[redacted]t]sharif[dot]edu
more ...
Homepage: http://sharif.ed[redacted]



Department of Physics › **Seyyed Mohamma[redacted]** › **Information**

Web Content Display

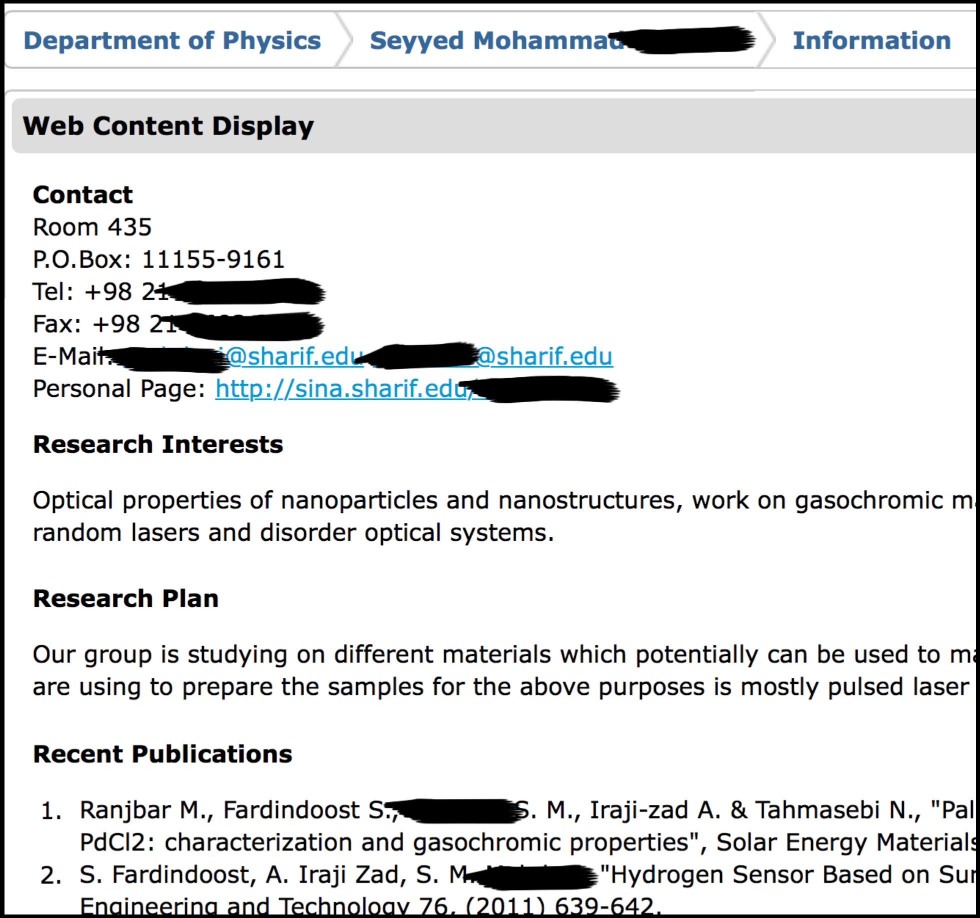
Contact
Room 435
P.O.Box: 11155-9161
Tel: +98 21[redacted]
Fax: +98 21[redacted]
E-Mai[redacted]@sharif.edu[redacted]@sharif.edu
Personal Page: http://sina.sharif.edu/[redacted]

Recent Publications

1. Ranjbar M., Fardindoost S.,[redacted]S. M., Iraji-zad A. & Tahmasebi N., "Pal PdCl2: characterization and gasochromic properties", Solar Energy Materials
2. S. Fardindoost, A. Iraji Zad, S. M[redacted] "Hydrogen Sensor Based on Su Engineering and Technology 76, (2011) 639-642.

Scenario Results

- Exploit the data available from
  - Web pages, publications, articles, etc.
- Produce a knowledge graph
  - Key people and connections
  - Technical capabilities and how they have changed over time



Web Content Display

Contact
Room 435
P.O.Box: 11155-9161
Tel: +98 21[redacted]
Fax: +98 21[redacted]
E-Mai[redacted]@sharif.edu[redacted]@sharif.edu
Personal Page: http://sina.sharif.edu/[redacted]

Research Interests

Optical properties of nanoparticles and nanostructures, work on gasochromic m
random lasers and disorder optical systems.

Research Plan

Our group is studying on different materials which potentially can be used to ma
are using to prepare the samples for the above purposes is mostly pulsed laser

Recent Publications

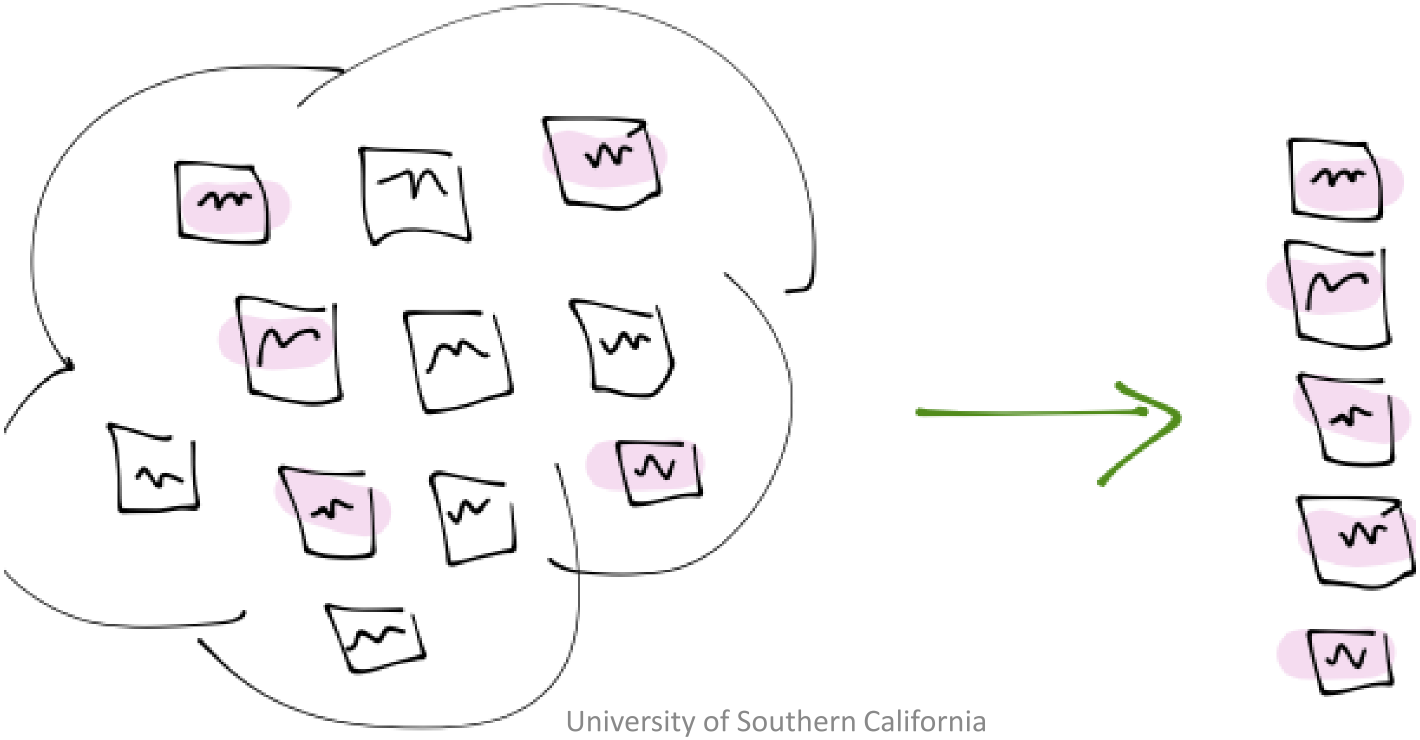
1. Ranjbar M., Fardindoost S.,[redacted]S. M., Iraji-zad A. & Tahmasebi N., "Pal
   PdCl2: characterization and gasochromic properties", Solar Energy Materials
2. S. Fardindoost, A. Iraji Zad, S. M[redacted]"Hydrogen Sensor Based on Sur
   Engineering and Technology 76, (2011) 639-642.

DIG Pipeline

- Crawling
- Extracting
- Cleaning
- Integration
- Computing simlarity
- Entity resolution
- Graph construction
- Query, analysis, and visualization

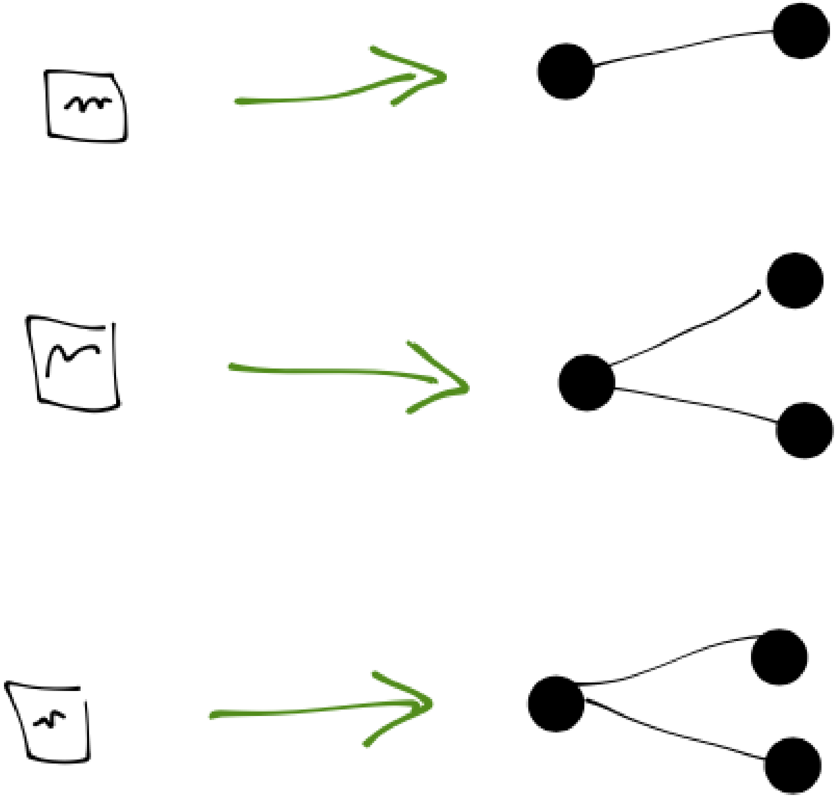
Crawling

- Challenge: how to crawl just the relevant pages
- Approach:
  - Uses the Apache Nutch framework for Web pages
  - Uses Karma to extract pages from the deep Web



Extracting

- Need to produce a structured representation for indexing and linking
- Highly configurable architecture for extractors
  - Learning of landmark extractors for structured data
  - Trainable CRF-based extractors for unstructured data
  - Uses Mechanical Turk to crowd source training data

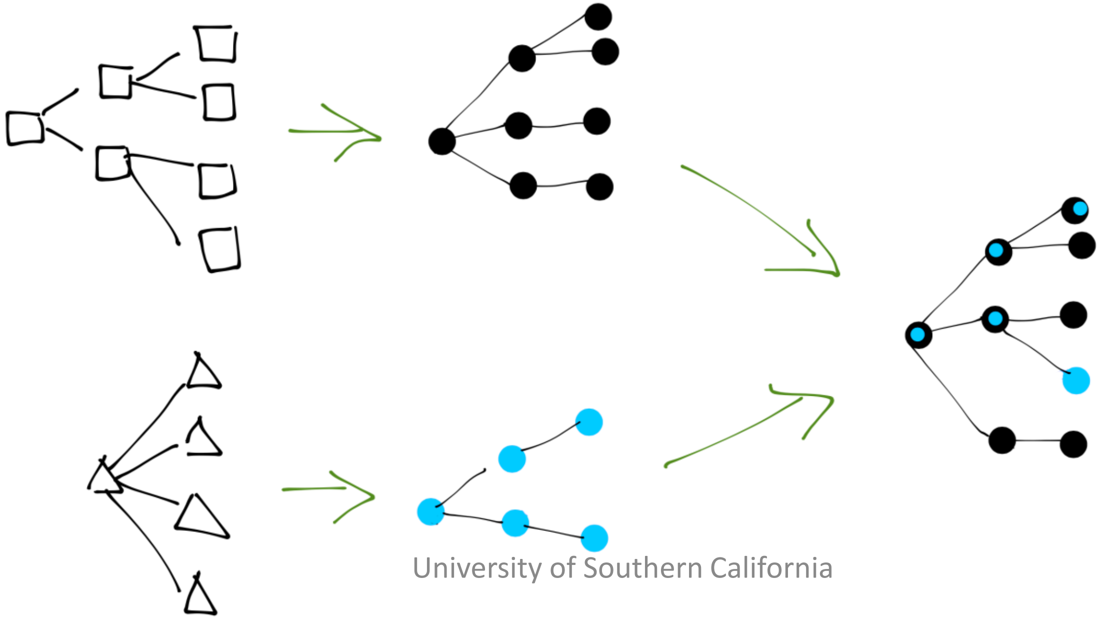


Cleaning

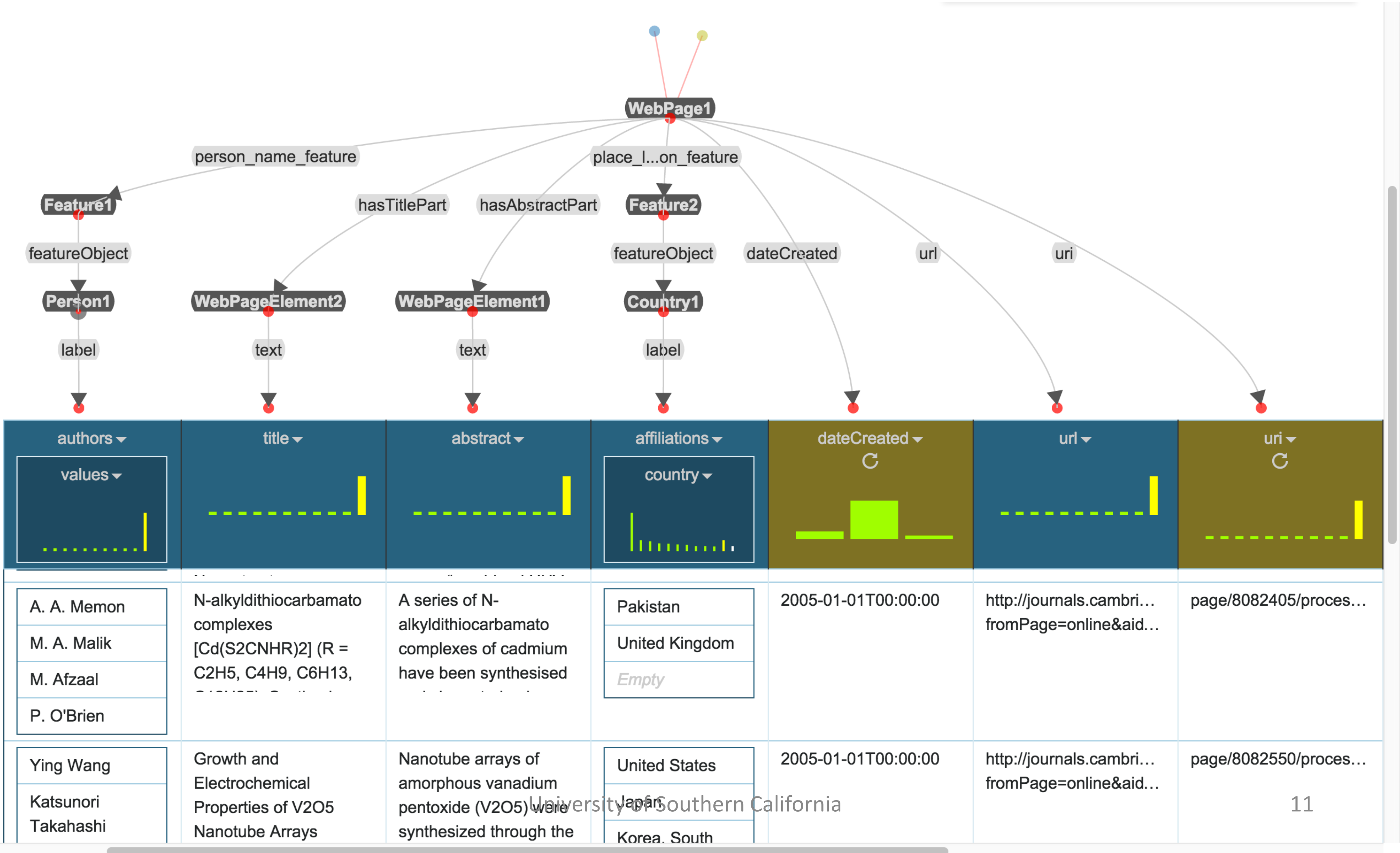
- Cleaning and normalization to support analysis and linking
  - – Visualization showing data distribution
  - – Learned transformations from examples
  - – Cleaning programs written in Python

Integration

- Need to align the data across extracted data and structured sources
- Performed using a data integration tool called Karma
    - Karma maps arbitrary sources into a shared domain vocabulary (schema alignment)
    - Uses machine learning to minimize user effort



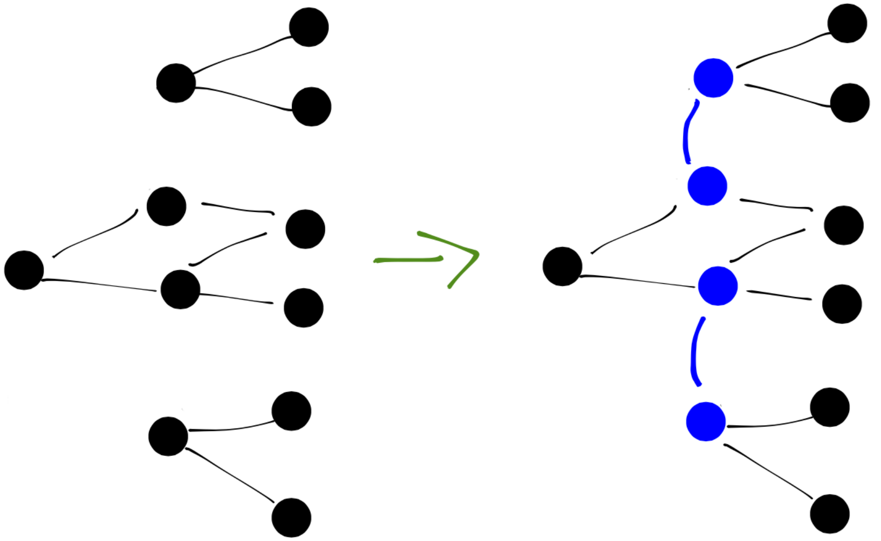
Integration Using Karma



authors ▾ values ▾	title ▾	abstract ▾	affiliations ▾ country ▾	dateCreated ▾	url ▾	uri ▾
A. A. Memon M. A. Malik M. Afzaal P. O'Brien	N-alkyldithiocarbamato complexes $[Cd(S_2CNHR)_2]$ (R = C_2H_5, C_4H_9, C_6H_{13},	A series of N-alkyldithiocarbamato complexes of cadmium have been synthesised	Pakistan United Kingdom Empty	2005-01-01T00:00:00	http://journals.cambri… fromPage=online&aid…	page/8082405/proces…
Ying Wang Katsunori Takahashi	Growth and Electrochemical Properties of V_2O_5 Nanotube Arrays	Nanotube arrays of amorphous vanadium pentoxide (V_2O_5) were synthesized through the	United States Japan Korea, South	2005-01-01T00:00:00	http://journals.cambri… fromPage=online&aid…	page/8082550/proces…

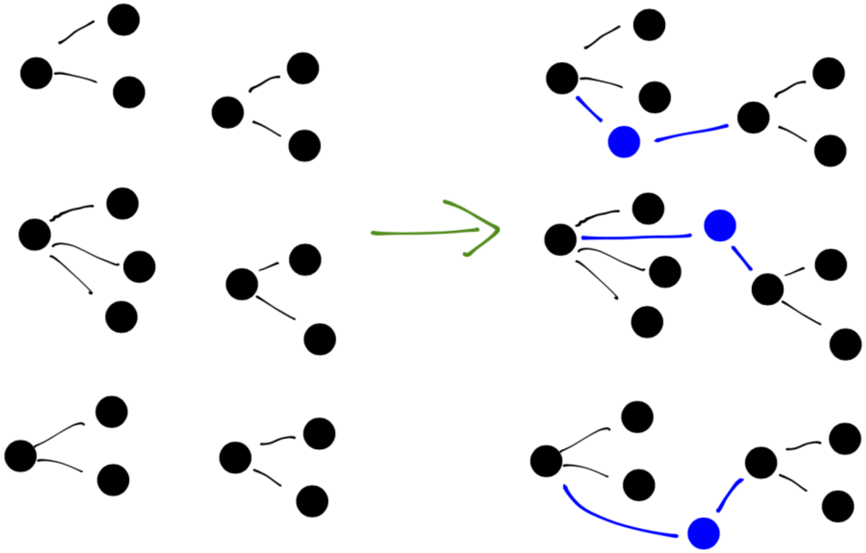
Similarity

- Computes similarity across text fields and images
  - Image similarity done using DeepSentiBank
  - Text similarity done using Minhash/LSH



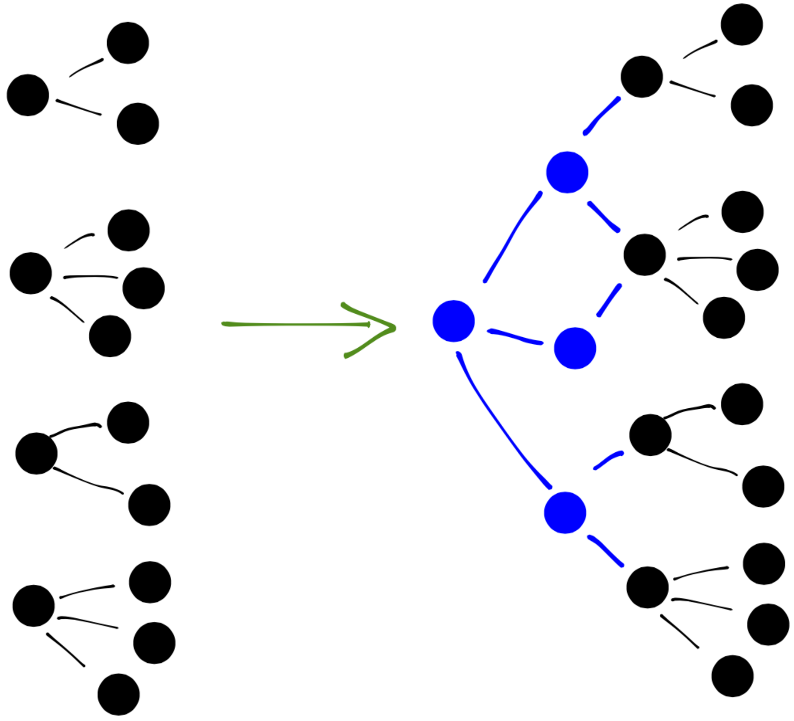
Entity Resolution

- Finds matching entities
- Reference source
  - Match against source to disambiguate entities
  - E.g., geonames for locations
- No reference source
  - Combine entities by considering the similarity across multiple fields



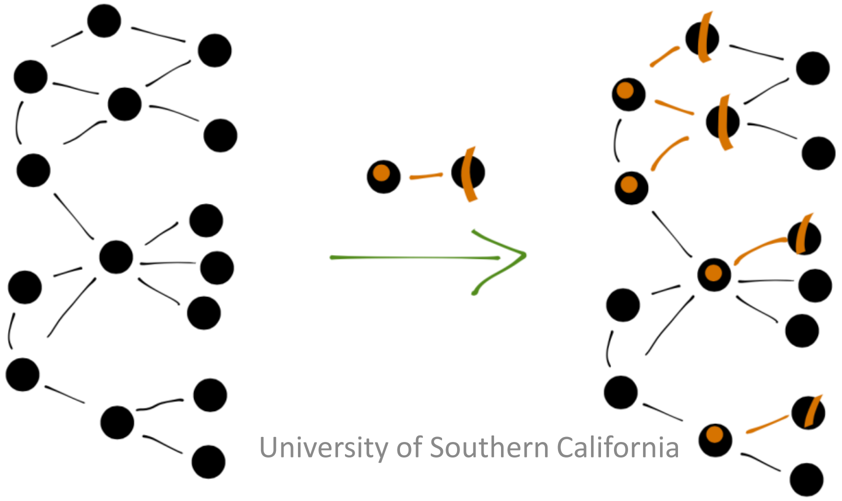
Graph Construction

- Data is integrated into a graph that can be queries and analyzed
  - Data stored in HDFS
  - Data represented in a common language JSON-LD
  - Represented using a common terminology



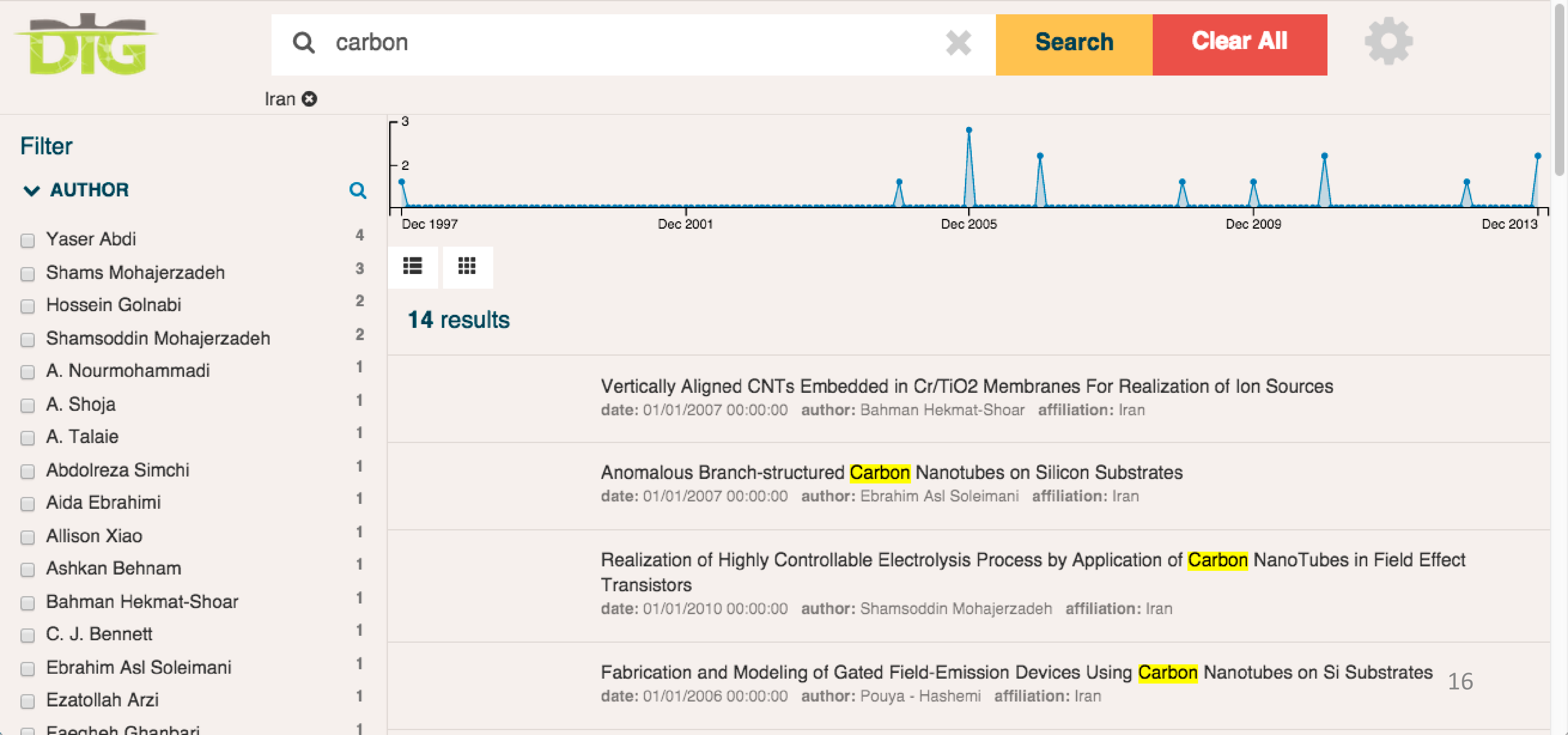
Query, Analysis and Visualization

- Challenge: support efficient querying against the graph
- Employ ElasticSearch to provide keyword querying, faceted browsing, and aggregation queries



Query, Analysis & Visualization

- Visualization interface that provides faceted queries, timeslines, maps, etc.



Filter

AUTHOR

□ Yaser Abdi	4
□ Shams Mohajerzadeh	3
□ Hossein Golnabi	2
□ Shamsoddin Mohajerzadeh	2
□ A. Nourmohammadi	1
□ A. Shoja	1
□ A. Talaie	1
□ Abdolreza Simchi	1
□ Aida Ebrahimi	1
□ Allison Xiao	1
□ Ashkan Behnam	1
□ Bahman Hekmat-Shoar	1
□ C. J. Bennett	1
□ Ebrahim Asl Soleimani	1
□ Ezatollah Arzi	1
□ Faegheh Ghanbari	1

14 results

Vertically Aligned CNTs Embedded in Cr/TiO2 Membranes For Realization of Ion Sources
date: 01/01/2007 00:00:00 **author:** Bahman Hekmat-Shoar **affiliation:** Iran

Anomalous Branch-structured Carbon Nanotubes on Silicon Substrates
date: 01/01/2007 00:00:00 **author:** Ebrahim Asl Soleimani **affiliation:** Iran

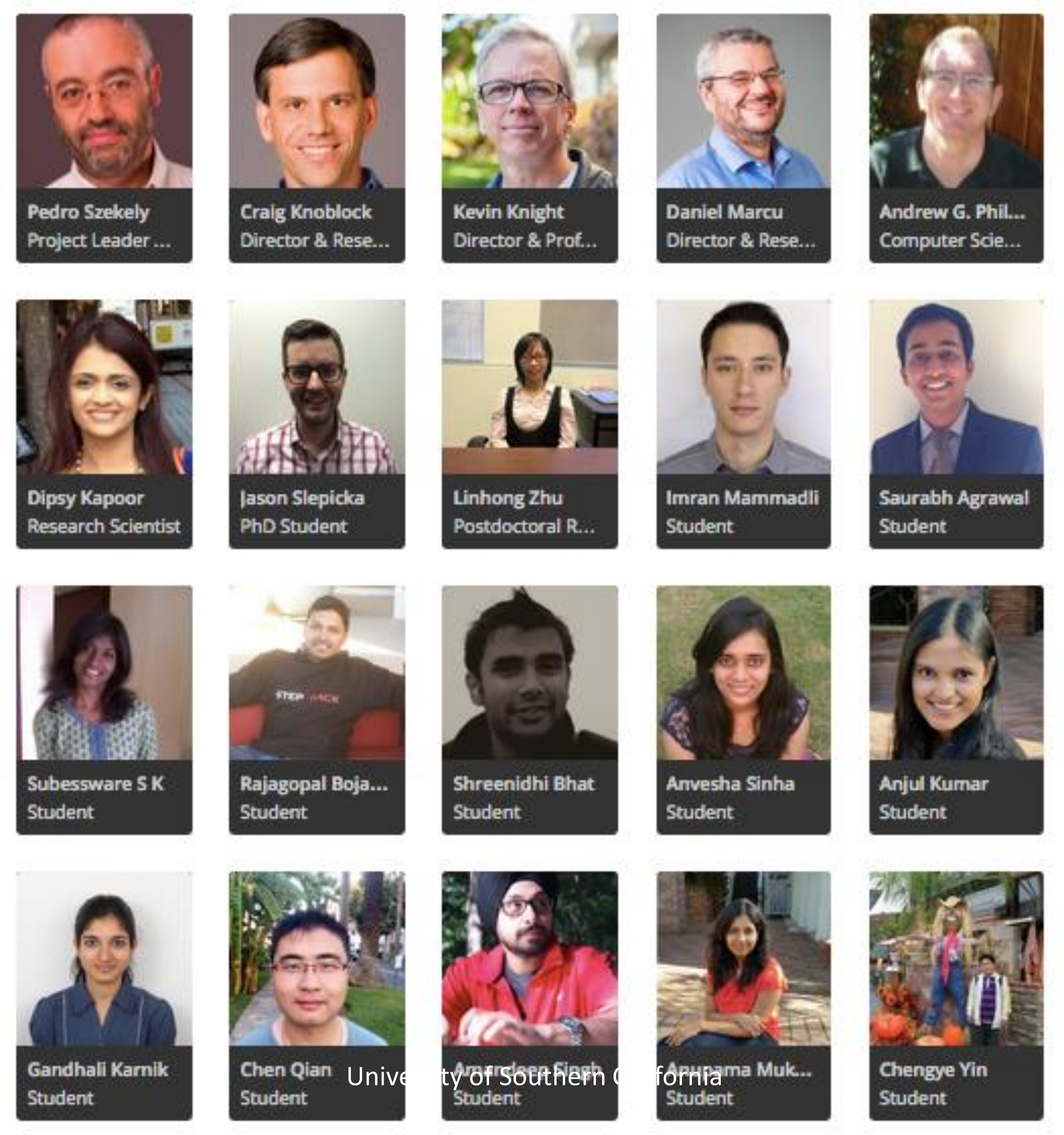
Realization of Highly Controllable Electrolysis Process by Application of Carbon NanoTubes in Field Effect Transistors
date: 01/01/2010 00:00:00 **author:** Shamsoddin Mohajerzadeh **affiliation:** Iran

Fabrication and Modeling of Gated Field-Emission Devices Using Carbon Nanotubes on Si Substrates
date: 01/01/2006 00:00:00 **author:** Pouya - Hashemi **affiliation:** Iran

Discussion

- Technology that can provide dramatic new insights from data that is already available
- Applies to a wide range of problems
  - Determining the nuclear know-how of a given country
    - Technologies, key scientists, relevant organizations
  - Combating human trafficking
  - Understanding trends in technical areas
    - E.g., Material Science
  - Analyzing the competitive landscape of companies
  - and many other domains with massive quantities of data

USC DIG Team



Pedro Szekely
Project Leader ...

Craig Knoblock
Director & Rese...

Kevin Knight
Director & Prof...

Daniel Marcu
Director & Rese...

Andrew G. Phil...
Computer Scie...

Dipsy Kapoor
Research Scientist

Jason Slepicka
PhD Student

Linhong Zhu
Postdoctoral R...

Imran Mammadli
Student

Saurabh Agrawal
Student

Subesswаre S K
Student

Rajagopal Boja...
Student

Shreenidhi Bhat
Student

Anvesha Sinha
Student

Anjul Kumar
Student

Gandhali Karnik
Student

Chen Qian
Student

Amandeep Singh
Student

Anupama Muk...
Student

Chengye Yin
Student

Acknowledgements

- Collaborators

InferLink
NEXT CENTURY CORPORATION
COLUMBIA UNIVERSITY
NASA
JPL

- Sponsor
  - DARPA
    - AFRL contract number FA8750-14-C-0240

Thanks!

- More information:
  - Homepage
    - isi.edu/~knoblock
  - DIG
    - usc-isi-i2.github.io/dig
  - Karma
    - usc-isi-i2.github.io/karma

DIG

Domain-specific Insight Graphs

Information Sciences Institute

USC University of Southern California