Hybrid Human-machine Systems

Lecture 5
Gianluca Demartini
University of Sheffield
Background

• Hybrid systems
  – Combining the scalability of machines and the quality of human intelligence
Hybrid Systems: Key Issues

• The role of machine (i.e., algorithm) and humans
  – use only humans? both? who’s doing what?
• Quality control
• Optimization: What to crowdsourcing
• Scalability: How much to crowdsourcing
Thinking About Hybrid Systems
Example: Hybrid Image Search

Yan, Kumar, Ganesan, CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones, Mobisys 2010.
Example: Hybrid Data Integration

<table>
<thead>
<tr>
<th>paper</th>
<th>conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data integration</td>
<td>VLDB-01</td>
</tr>
<tr>
<td>Data mining</td>
<td>SIGMOD-02</td>
</tr>
</tbody>
</table>

- **Generate plausible matches**
  - paper = title, paper = author, paper = email, paper = venue
  - conf = title, conf = author, conf = email, conf = venue

- **Ask users to verify**

Does attribute paper match attribute author?

<table>
<thead>
<tr>
<th>paper</th>
<th>conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data integration</td>
<td>VLDB-01</td>
</tr>
<tr>
<td>Data mining</td>
<td>SIGMOD-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>email</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLAP</td>
<td>Mike</td>
<td>mike@a</td>
<td>ICDE-02</td>
</tr>
<tr>
<td>Social media</td>
<td>Jane</td>
<td>jane@b</td>
<td>PODS-05</td>
</tr>
</tbody>
</table>

McCann, Shen, Doan: Matching Schemas in Online Communities. ICDE, 2008
Example: Hybrid Query Processing

Use the crowd to answer DB-hard queries

Where to use the crowd:
• Find missing data
• Make subjective comparisons
• Recognize patterns

But not:
• Anything the computer already does well

Crowdsourced Data Management Applications

- Relational
  - information extraction
  - schema matching
  - entity resolution
  - building structured KBs
  - sorting
  - top-k
  - ...

- Beyond relational
  - graph search
  - classification
  - mobile image search
  - social media analysis
  - question answering
  - NLP
  - text summarization
  - sentiment analysis
  - semantic wikis
  - ...

Qurk (MIT)

- **Goal:** crowd-source comparisons, missing data
- **Basis:** SQL3 + UDF
  - UDF encapsulate crowd input
  - special template language for crowd UDFs
  - specify UI, quality control, ... (possibly opt. hints)
Qurk Example [Markus et al. CrowdCrowd 2011]

• Task: Find all women in a “people” database

• Schema

  CREATE TABLE people(
    name varchar(256),
    photo blob);

• Query

  SELECT name
  FROM people p
  WHERE isFemale(p);
Qurk Example [Markus et al. CrowdCrowd 2011]

• Task: Find all women in a “people” database

• Schema

  CREATE TABLE people(
    name varchar(256),
    photo blob );

• Query

  SELECT name
  FROM people p
  WHERE isFemale(p);

  TASK isFemale(tuple) TYPE: Filter
  Question: “is %s Female”,
            tuple[“photo”]

  YesText: “Yes”
  NoText: “No”
Qurk Example [Markus et al. Crowd Crowd 2011]

Is ___ Female?

 TASK isFemale(tuple) TYPE: Filter
 Question: “is %s Female”,
    tuple[“photo”]

 YesText: “Yes”
 NoText: “No”
The magic is in the Templates

• Templates generate UIs for different kinds of crowd-sourcing tasks
  – filters: Yes / No questions
  – joins: comparisons between two tuples (equality)
  – order by: comparisons between two tuples (gt?)
  – generative: crowd-source attribute values

• Templates also specify quality control; e.g.,
  COMBINER: MajorityVote
Crowdsourcing DB Systems

• Fundamentally new way of tackling data management issues using large networks of anonymous users

• At this point, first interesting systems and results, but still more questions than answers
  – Hot research topic

• Unique, unexpected issues
  – “My database hates me”
Problem: Populate Infobboxes
Solution: IE Using Machine + Human

Hemingway was an American author ...

The American readers ...

Infobox

Born    July 21, 1899
Nationality    American

Train a “nationality” extractor

Apply extractor to new pages to extract nationalities.

Verify with crowd
Recruiting Model: Search Advertising

- Interrupt user in the middle of a primary task – searching for information on Ray Bradbury
- Ask if user is willing to contribute

- Evaluate different UIs: pop-up, highlight, icon – in terms of intrusiveness and willingness to contribute
Ray Bradbury

From Wikipedia, the free encyclopedia

Ray Douglas Bradbury (born August 22, 1920) is an American literary, fantasy, horror, science fiction, and mystery writer best known for *The Martian Chronicles*, a 1950 book which has been described both as a short story collection and a novel, and his 1953 dystopian novel *Fahrenheit 451*. He is widely considered to be one of the greatest and most popular American writers of speculative fiction during the twentieth century.

Contents [show]

Beginnings

Bradbury was born in Waukegan, Illinois, to a Swedish immigrant father and telephone lineman. His paternal grandfather and uncle were newspaper publishers.[1] Bradbury was a reader and writer throughout his youth in Waukegan. He used this library as a setting for much of his novel *Something Wicked This Way Comes*, and depicted Waukegan as “Green Town” in some of his other semi-autobiographical novels — *Dandelion Wine*, *Farewell Summer* — as well as in many of his short stories.[2]

He attributes his lifelong habit of writing every day to an incident in 1932 when a carnival entertainer, Mr. Electrico,[4] touched him with an electrified sword, made his hair stand on end, and shouted, “Live forever!”

The Bradbury family lived in Tucson, Arizona, in 1926–27 and 1932–33 as his father pursued employment, each time returning to Waukegan, and eventually settled in Los Angeles in 1934, when Ray was thirteen.

Bradbury graduated from the Los Angeles High School in 1938 but chose not to attend college. Instead, he sold newspapers at the corner of South Notton Avenue and Olympic Boulevard. He continued to educate himself at the local library, and having been influenced by science fiction heroes like *Flash Gordon* and *Buck Rogers*, he began to publish science fiction stories in *fanzines* in 1938. Ray was invited by Forrest J Ackerman to attend the now legendary Clifton's Cafeteria Science Fiction Club. Here Ray met the writers Robert A. Heinlein, Emil Petaia, Fredric Brown, Henry Kuttner, Leigh Brackett, and Jack Williamson.
Ray Bradbury
From Wikipedia, the free encyclopedia

Ray Douglas Bradbury (born August 22, 1920) is an American literary, fantasy, horror, science fiction, and mystery writer best known for The Martian Chronicles, a 1950 book which has been described both as a short story collection and a novel, and his 1953 dystopian novel Fahrenheit 451. He is widely considered to be one of the greatest and most popular American writers of speculative fiction during the twentieth century.

Beginnings

Bradbury was born in Waukegan, Illinois, and telephone lineman. His paternal grandparents had moved to Waukegan from Scotland. Bradbury was a reader and writer throughout childhood. Waukegan was depicted in stories such as Dandelion Wine, Farewell Summer and Fireball. He attributes his lifelong habit of writing to his father, Mr. Electrico,[4] who touched him with an electric wire and told him he would burn "forever." The Bradbury family lived in Tucson, Arizona, where Ray's father worked for the employment, each time returning to Waukegan. As Ray was thirteen.

Bradbury graduated from the Los Angeles High School in 1938 but chose not to go to college. Instead, he sold newspapers at the corner of South Norton Avenue and Olympic Boulevard. He continued to educate himself at the local library, and having been influenced by science fiction heroes like Flash Gordon and Buck Rogers, he began to publish science fiction stories in fanzines in 1938. Ray was invited by Forrest J Ackerman to attend the now legendary Clifton's Cafeteria Science Fiction Club. Here Ray met the writers Robert A. Heinlein, Emil Petaja, Fredric Brown, Henry Kuttner, Leigh Brackett, and Jack Williamson...
Other examples of hybrid systems
Facebook is not waiting for its initial public offering to make its first big purchase.

In its largest acquisition to date, the social network has purchased Instagram, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.
ZenCrowd

• Combine both algorithmic and manual linking
• Automate manual linking via crowdsourcing
• Dynamically assess human workers with a probabilistic reasoning framework
ZenCrowd Architecture

Entity Factor Graphs

• Graph components
  – Workers, links, clicks
  – Prior probabilities
  – Link Factors
  – Constraints

• Probabilistic Inference
  – Select all links with posterior prob >τ

2 workers, 6 clicks, 3 candidate links
Experimental Evaluation

• Datasets
  – 25 news articles from
    • CNN.com (Global news)
    • NYTimes.com (Global news)
    • Washington-post.com (US local news)
    • Timesofindia.indiatimes.com (India news)
    • Swissinfo.com (Switzerland local news)
  – 40M entities (Freebase, DBPedia, Geonames, NYT)

<table>
<thead>
<tr>
<th></th>
<th>US Workers</th>
<th></th>
<th>Indian Workers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>A</td>
<td>P</td>
</tr>
<tr>
<td>GL News</td>
<td>0.84</td>
<td>0.87</td>
<td>0.90</td>
<td>0.67</td>
</tr>
<tr>
<td>US News</td>
<td>0.64</td>
<td>0.68</td>
<td>0.78</td>
<td>0.55</td>
</tr>
<tr>
<td>IN News</td>
<td>0.84</td>
<td>0.82</td>
<td>0.89</td>
<td>0.75</td>
</tr>
<tr>
<td>SW News</td>
<td>0.72</td>
<td>0.80</td>
<td>0.85</td>
<td>0.61</td>
</tr>
<tr>
<td>All News</td>
<td>0.80</td>
<td>0.81</td>
<td>0.88</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Worker Selection

![Graph showing worker precision vs. number of tasks for US and IN workers.](image)

![Graph showing precision vs. top K workers.](image)
Lessons Learnt

• Crowdsourcing + Prob reasoning works!

• But
  – Different worker communities perform differently
  – Many low quality workers
  – Completion time may vary (based on reward)

• Need to find the right workers for your task
  (see WWW13 paper)
ZenCrowd Summary

• ZenCrowd: Probabilistic reasoning over automatic and crowdsourcing methods for entity linking
• Standard crowdsourcing improves 6% over automatic
• 4% - 35% improvement over standard crowdsourcing
• 14% average improvement over automatic approaches

http://exascale.info/zencrowd/

• Follow up-work (VLDBJ):
  – Also used for instance matching across datasets
  – 3-way blocking with the crowd
**Blocking for Instance Matching**

- Find the instances about the same real-world entity within two datasets
- Avoid Comparison of all possible pairs
  - Step 1: cluster similar items using a cheap similarity measure
  - Step 2: n*n comparison within the clusters with an expensive measure
3-steps Blocking with the Crowd

- Crowdsourcing as the most expensive similarity measure
CrowdQ – Crowd-powered Query Understanding
birthdate of the mayor of the capital city of Italy
capital city of Italy

Rome
Italy, Capital
Marino ran the 2013 election for Mayor of Rome with the support of a centre-left alliance. After leading in the first round he was elected (on 10 June) Mayor of Rome at the second ballot, winning 63.9% of the votes in a run-off against the centre-right candidate, the outgoing mayor Gianni Alemanno.

Ignazio Marino - Wikipedia, the free encyclopedia

Mayor of Rome - Wikipedia, the free encyclopedia

The Mayor of Rome (Italian: Sindaco di Roma Capitale) is an elected politician who serves as the Mayor of the city of Rome, Italy.
March 10, 1955 (age 59 years)

Ignazio Marino, Date of birth
Motivation

• Web Search Engines can answer simple factual queries directly on the result page

• Users with complex information needs are often unsatisfied

• Purely automatic techniques are not enough

• We want to solve it with Crowdsourcing!
CrowdQ

• CrowdQ is the first system that uses crowdsourcing to
  – *Understand* the intended meaning
  – *Build* a structured query template
  – *Answer* the query over Linked Open Data

<table>
<thead>
<tr>
<th>City</th>
<th>Mayor</th>
<th>Birthdate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rome, Italy</td>
<td>Gianni Alemanno</td>
<td>March 3, 1958</td>
</tr>
<tr>
<td>Venice, Italy</td>
<td>Giorgio Orsoni</td>
<td>August 29, 1946</td>
</tr>
<tr>
<td>Milan, Italy</td>
<td>Giuliano Pisapia</td>
<td>May 20, 1949</td>
</tr>
</tbody>
</table>

**Cities in Italy** | **Italy** Travel Guide

Learn about the best cities in Italy to visit, and some Italian cities you might never have heard of before. These cities in Italy are all great for visitors.

**Top Ten Cities for Visitors to Italy** - Top Italian Cities to See
goitaly.about.com/od/planningandinformation/tp/topcities.htm

Italy has many beautiful and historic cities that are well worth a visit. Here are our picks for the ten best cities for visitors to Italy.

**Italian Cities and Towns - Italy**
en.comuni-italiani.it/

Information and statistics on Italian Regions, Provinces, and Municipalities. All Cities.
CrowdQ Architecture

**Off-line**: query template generation with the help of the crowd

**On-line**: query template matching using NLP and search over open data
Hybrid Human-Machine Pipeline

Q= birthdate of actors of forrest gump

Query annotation

Noun

Noun

Named entity

Verification

Is forrest gump this entity in the query?

Entity Relations

Which is the relation between: actors and forrest gump starring

Schema element

Starring <dbpedia-owl:starring>

Verification

Is the relation between:
Indiana Jones – Harrison Ford
Back to the Future – Michael J. Fox
of the same type as
Forrest Gump - actors
Structured query generation

Q= birthdate of actors of forrest gump

SELECT ?y ?x
  ?z <rdfs:label> ‘Forrest Gump’ }
## Overview of hybrid systems

<table>
<thead>
<tr>
<th>Year</th>
<th>Cit.</th>
<th>Domain</th>
<th>Data Type</th>
<th>Human role</th>
<th>Incentive</th>
<th>Time constrains</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>[52]</td>
<td>Database</td>
<td>Graph</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2012</td>
<td>[56]</td>
<td>Data Integration</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>[34]</td>
<td>Info Extraction</td>
<td>Unstruct. text</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>[48]</td>
<td>Biomedical</td>
<td>Ontology</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>[43]</td>
<td>Personal assistance</td>
<td>Unstruct. text</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2013</td>
<td>[27]</td>
<td>Biomedical</td>
<td>Unstruct. text</td>
<td>Post-p.</td>
<td>Fun</td>
<td>Batch</td>
</tr>
<tr>
<td>2014</td>
<td>[53]</td>
<td>Search</td>
<td>Image</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
</tbody>
</table>
Overview of hybrid systems

• Balance between systems that use the human component as pre-processing or post-processing of data (11 vs 13)
• Mostly monetary reward
• Majority of systems perform batch data processing rather than real-time jobs
• In 2014 we can observe a decreased number of hybrid human-machine systems being propose: focus on solving core problems rather than building new systems
Summary

• Crowdsourcing big data can make you go bankrupt! -> hybrid systems
• When ask a human, when trust the machine
• Hybrid (human in the loop)
  – Pre-processing: training data for ML
  – Post-processing: based on confidence scores
  – Mix: active learning