From Information Extraction Research to Vertical Search Products: The Semantic Gap is More Than the Structure Divide

Kevin C. Chang
University of Illinois at Urbana-Champaign
Mass Information Extraction: Building Deep Vertical Search Engines

Kevin C. Chang
University of Illinois at Urbana-Champaign
Why professor is not fit for start-up, and why he should do it anyway.

Kevin C. Chang
Cazoodle Inc.
The Pretext
How do you greet people in your culture?

What have you been searching lately?
What have you been searching lately?

- The university and areas of Kevin Chang?
- The email of Jiawei Han?
- Customer service phone number of Amazon?
- What profs are doing databases at UIUC?
- The papers and presentations of WWW 2013?
- Due date of WWW 2014?
- Sale price of “Canon PowerShot A400”?
- “Hamlet” books available at bookstores?
Data of all sorts--- Prevalent on the Web!

Kevin Chen-Chuan Chang

<table>
<thead>
<tr>
<th>Departing</th>
<th>Flight</th>
<th>City</th>
<th>Date &amp; Time</th>
<th>Arriving</th>
<th>Date &amp; Time</th>
<th>Aircraft Type</th>
<th>Cabin</th>
<th>Flight Miles</th>
<th>Meals</th>
<th>Travel Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMERICAN AIRLINES OPERATED BY AMERICAN EAGLE</td>
<td>440</td>
<td>CMI</td>
<td>07/26/2006 08:24 AM</td>
<td>ORD</td>
<td>07/30/2006 09:14 AM</td>
<td>ACR</td>
<td>Economy</td>
<td>125</td>
<td>$79</td>
<td>7 hr 4 min</td>
</tr>
<tr>
<td>AMERICAN AIRLINES</td>
<td>1550</td>
<td>ORD</td>
<td>07/26/2006 12:44 PM</td>
<td>SFO</td>
<td>07/30/2006 09:30 PM</td>
<td>M32</td>
<td>Economy</td>
<td>1540</td>
<td>$79</td>
<td>10 hr 44 min</td>
</tr>
</tbody>
</table>

Excerpt from Page 1: "... This chapter introduces fundamental concepts and definitions regarding databases, including properties common to databases, prevalent database models, a brief ..."
The Web is a Huge Supermarket!
What you search is what you want?
Our View—
Web is “interlinked pages” of “data objects”.

[Cazoodle logo]
Challenges on the Web come in “dual”:
Getting **access** to the **structured** information!

**Kevin’s 4-quadrents:**

<table>
<thead>
<tr>
<th>Access</th>
<th>Deep Web</th>
<th>Surface Web</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structure</strong></td>
<td><img src="image" alt="MetaQuerier" /></td>
<td><img src="image" alt="WISDM" /></td>
</tr>
</tbody>
</table>

**WISDM**
Web Indexing and Search for Data Mining
The Beginning
Core challenge of search:

*Bridging the Semantic Gap*

What does user want?

What does the Web present?
Information Extraction Research

Settings
- Named entity tagging
- Relation extraction
- Event extraction
- Record extraction

Techniques
- Textual Patterns: Dictionary, Regular Expression
- Statistical Models: HMM, CRF.
- Page Templates: Wrapper Induction.
Vertical Search Products
The view is incomplete and therefore limited.

From Information Extraction to Vertical Search: How do they connect?
In 2004, Homestore.com (now Move.com) called me.

- They were king in Real Estate for Sale (powering up Realtor.com).
- Wanted to expand in two ways—
  - Real Estate for Sale in Canada and UK.
  - Apartments for Rent in US.
- But were stuck in either way.

- Why?
From Information Extraction to Vertical Search:

... ...

real vertical search engines are rare.
They do not connect, for the most part.

How does the industry work?

- **By standards/protocols**: Near-standardized DB.
  - RealEstate@US: MLS, National Association of Realtors since 1800’s. (cf. realtor.com)

- **By advertisement**: Publisher feeds to advertise.
  - Online Shopping: Google Products.

- **By hands** (desperately) – Guess who?
The only exception is Job Search.
And, they are rather successful.
The Pain.
Students, how do you find apartments to rent?

Mom and dad looked everywhere in the world.

Now you look everywhere on the Web!
Professors, how do you find grants to beg?
There is even a "feed management" industry.
The Solutions
Cazoodle’s Mission: Structured data-aware Search
Cazoodle’s Technology:
Web-scale Data Integration

<table>
<thead>
<tr>
<th>Address</th>
<th>Rent</th>
<th>Area</th>
<th>Bedroom</th>
</tr>
</thead>
<tbody>
<tr>
<td>4916 W. Cornelia, 60641</td>
<td>$950</td>
<td>Portage Park</td>
<td>2BR</td>
</tr>
<tr>
<td></td>
<td>$995</td>
<td>Uptown/Buena Park</td>
<td>Studio</td>
</tr>
<tr>
<td>7501 N Eastlake Terrace, 60626</td>
<td>$710</td>
<td>Rogers Park</td>
<td>1BR</td>
</tr>
<tr>
<td>N LaSalle St at W Elm St, 60610</td>
<td>$1,300</td>
<td>Gold Coast/Streeter</td>
<td>1BR</td>
</tr>
<tr>
<td>N Clark at W Belmont, 60657</td>
<td>$900</td>
<td>Lakeview</td>
<td>2BR</td>
</tr>
</tbody>
</table>
Apartment Search: Find all apartments on the same map.
The Technology: Mass IE for Deep Search
Behind the Scene: Building Vertical Search

- **Source Discovery**
  - Find relevant sources.

- **Object-centric Crawling**
  - Navigate/query source to gather data.

- **Data Extraction**
  - Extract attributes and values from pages crawled.
IE for Vertical Search: Key Requirements.

- **Automation:**
  - Crawling + Extraction.

- **Usability:**
  - **Accuracy** to allow object-level search/filtering.

- **Scalability:** **Robustness** and **cost-effectiveness**
#1. IE needs to be contextual.

*It may not start with “given a set of result pages.”*
#2. IE should not stop at 70% accuracy.

Imagine price or address is 70% correct.
#3. IE shall be scalable to (more than) thousands of sources.

Resist the 80-20 law when it comes to vertical search.
#4. IE should be programmed to last.

Do not use programmers or graduate students.
#5. IE shall leverage human in the loop.

At least for the final step of quality assurance.
IE for Vertical Search: Data Agents

- **Modeling**: Activeness, a state machine.
- **Description**: High-level features.
- **Creation**: Non-programmers + Algorithms.
The Results
Technical-wise:

- Ability to
  - Scale beyond 10000 sources, with small agent team.
  - Work across domains and languages.

- Vertical search in various domains:

- Data factory for serving data to customers in various domains and countries.
Business-wise:

- **PlaceofMine.com**
  - 500K users/month, by advertisement revenue.

- **GrantForwrad.com**
  - 150 university/institution subscribers.

- **Data Factory**
  - PriceGrabber, Zap (Brazil), Comcast.
  - And many startups you do not want to entertain…
The Lessons, or The *Semantic Gaps*. 
#1. Professor:
Are you in the silicon valley?
Silicon Valley and the Prairie are different.

- When he put the Google paper in the qual exam, the comment was …
- And, he is evaluated by papers, proposals, citations, students graduated, and not …

- But things things have changed quickly at UIUC!
#2. Professor:

Do you start with a product?
Product and application are different.

- He often wrote intro of papers – “there are many applications…”, but which one?
- He looked at data factory, was amazed to see the huge demands– from startups.
- He looked at various vertical search domains.
#3. Professor:
Do you know the industry?
Industry and academic are different.

- Did he know how vertical services were built?
- The model may be quite disruptive which also means more barriers to cross.
  - They get paid to get data, while you spend money to crawl.
  - They use revenue to buy traffic, and use traffic to market.
#4. Professor: Do you like to beg?
VC and NSF are different.

- Even forced and trying, how many VCs has he entertained?
- Turns out it’s possible to bootstrap; just slower…
- And, going slow may be an advantage!!
#5. Professor: Do you write JavaScript?
Building products and doing experiments are different.

- He liked to delegate…

- There are some fields that require professors to practice two days a week. We should do that.
- Now he is CEO + System Admin + Janitor.
- And, the professor is good in JavaScript now.
#6. Professor:
Do you know anyone who doesn’t write a thesis?
People and graduate students are different.

- He thought every “student” is laboring for getting thesis done, and he can call him at 2am.

- After 10 quits he would know that you don’t manage people with only science.
#7. Professor: Do you design?
Designing product and designing an algorithm are different.

- He designed a lot of algorithms.
- But that is the part users do not see and care.

- He may find a domain where products are well defined— and technology fits right in.
- Or, he will realize that he needs to learn products— not too hard if you can design algorithms.
#8. Professor: Do you know users?
Users and reviewers are different.

He knew reviewers, but not so much about users.

Reviewers are brutal but they tell what they want.

Users—

- Do you know what they are doing?
- Do they know what they want?
#9. Professor:  
Do you know the market?
Products awareness and paper citations are different.

- He was used to a market that actively “pull” papers— the concept of marketing is minimum.
- In reality, things are just the opposite.
#10. Professor: Are you inspired by Google?
Every startup is different.

- His case might not be like Google.
  - He needs innovation in product forms.
  - He needs to deal with “vertical” and thus “segmented” market.

- Even worse, Google might not like vertical search.
And, Why He Should Do It, Anyway!
The Why’s are rather simple…

- Prescription for (near) mid-age crisis.
  - What was this paper about?

- Exit for the feels-like Ponzi Scheme.
  - Bring research to see the light of the world.

- Inspiration for research.
  - See problems with a more real perspective.

- Live a little deeper
  - Have some stories to tell.
  - Have fun and think big.
Thank You!