Weakly Supervised User Profile Extraction from Twitter

Jiwei Li, Alan Ritter and Eduard Hovy

School of Computer Science, Carnegie Mellon University

June 22nd, 2014
User Profile

- Work
  - Occupation: Actor
- Basic Info
  - Location: London, United Kingdom
  - Timeline:
    - 01/2011: Joined Facebook
  - Awards:
    - Male Youth Discovery of the Year Women's Press Club - 2001
    - Best Male Performance (Nationals 2005)
    - Favorite Leading Actor in a Broad Facade Broadway's Audience Award
    - Favorite Breakthrough Performer Audience Award - 2008
    - Favorite Actor in a Broadway Play Audience Award - 2011
    - Favorite Male Group (shared with a partner)
    - Broadway Auditions - 2012
    - Best Male (MTV Movie Awards - 2012)

- Gender
- Education
- Religion
- Spouse/Lover
- Job
- Interest
- Favorite Movie
- Favorite Food
Main Contribution

Automatic extraction of attributes from Twitter:
- Spouse
- Education
- Job
Main Contribution

- We use Google plus as distant supervision for user attribute extraction.
- We present a large-scale dataset for this task.
- We demonstrate the benefit of jointly reasoning about users’ social network structure.
Outline

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion
User Profile Extraction

**Work and Education**

- **Stanford University**
  Professor · Stanford, California · Sep 2009 to present

- **Cornell University**
  Post Doc · Ithaca, New York · Sep 2008 to Sep 2009
  See All Employers

- **Cornell University**
  Ithaca, New York

- **Carnegie Mellon University**
  PhD in Machine Learning · Machine Learning, AI · Computer Science · Pittsburgh, Pennsylvania

- **University of Ljubljana**
  faculty of computer and information science · Ljubljana, Slovenia

- **Gimnazija Bezigrad**
  Class of 1999 · Ljubljana, Slovenia

**Places Lived**

- **Palo Alto, California**
  Current City

- **Sentjost nad Horjulom**
  Hometown

**Basic Information**

- Religious Views: Catholic

**Contact Information**

- Mobile: (650) 725-3711
- Phones:
- Address: 353 Serra Mall
  Stanford, CA, United States 94305
- Website: http://cs.stanford.edu/~jure
- Facebook: http://facebook.com/jure.leskovec
Motivation/Introduction

Fernando Pereira

Works at Google
Attended University of Lisbon
Lives in Palo Alto, California

3,834 have him in circles · 10 in common

Occupation
Distinguished researcher

Employment

Google
Research director, 2008 - present

LNEC, Lisbon
1975 - 1977

SRI International
1982 - 1989

AT&T
1989 - 2000

University of Pennsylvania
2001 - 2007

Education

University of Lisbon
Mathematics, 1969 - 1975

University of Edinburgh
Artificial intelligence, 1977 - 1982

Basic Information

Gender
Male

Relationship
Married
Why Profile Extraction?
Why Profile Extraction?

- Friend Recommendation
Why Profile Extraction?

- Friend Recommendation
- Target Advertising (Movie, Book ... )
Motivation/Introduction

Why Profile Extraction?

- Friend Recommendation
- Target Advertising (Movie, Book ... )

Already in a relationship with someone?
User Profile Extraction

Bill Simmons
- Columnist for The Ringer, NBA analyst, author of "The Book of Basketball"
- Twitter: @BillSimmons
- Description: Bill Simmons has more power/equity/say than anyone else in the NBA even Popovich, Doc, Spo + 30G! Flip Saunders!!! WTF????

Kenny Smith
- Basketball analyst for TNT
- Twitter: @TheJetOnTNT
- Description: championships and Emmy's, Kenny Smith

The Universe - kennythetjetsmith.com

Followed by NBA.
Twitter serves as a wonderful source:
Twitter serves as a wonderful source:

- **Text Level Evidence**

  - Omar Rasheed
    - HARVARD ACCEPTED ME😭😭😄
    - pic.twitter.com/VPsRo97VzE
  - Ana García Puyol
    - It’s been an honor studying at @Harvard for two years and going to class with really talented people who I can now call friends.
    - @Harvard
  - Ana García Puyol
    - Think! #harvard14 #graduation @Harvard
    - University instagram.com/p/olfB_ljCEg/
Motivation/Introduction

Twitter serves as a wonderful source:

- Text Level Evidence
- Network Information
Twitter serves as a wonderful source:

- **Text Level Evidence**
- **Network Information**
  - Homophily: People sharing more attributes have a higher chance of becoming friends in social media
Question

Unstructured Twitter data → Structured User Profile?
Related Work
Related Work

User Attribute Extraction/Identification

Gender (Ciot et al., 2013; Liu and Ruths, 2013)
Age (Rao et al., 2010)
Political Polarity (Pennacchiotti et al., 2011)
Relying on external political websites
Related Work

User Attribute Extraction/ Identification

- Gender (Ciot et al., 2013; Liu and Ruths, 2013)
Related Work

User Attribute Extraction/ Identification

- Gender (Ciot et al., 2013; Liu and Ruths, 2013)
- Age (Rao et al., 2010)
Related Work

User Attribute Extraction/ Identification

- Gender (Ciot et al., 2013; Liu and Ruths, 2013)
- Age (Rao et al., 2010)
  - Relying on Amazon Mechanical Turk
Related Work

User Attribute Extraction/ Identification

- **Gender** (Ciot et al., 2013; Liu and Roberts, 2013)
- **Age** (Rao et al., 2010)
  - Relying on Amazon Mechanical Turk
- **Political Polarity** (Pennacchiotti et al., 2011)
  - Relying on external political websites
Dataset Creation

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion
Dataset Creation

Challenge:

- Lack of Training Data
Dataset Creation

Distant Supervision
Dataset Creation

Distant Supervision

Paris is the capital and most populous city of France. The capital of France is Paris.
Distant Supervision

- Relation Extraction (Mintz et al., 2009)
Distant Supervision

- Relation Extraction (Mintz et al., 2009)

IsCapital (Paris, France)
IsCapital (London, Britain)
Distant Supervision

- Relation Extraction (Mintz et al., 2009)
**Distant Supervision**

- Relation Extraction (Mintz et al., 2009)

- Paris is the capital and most populous city of France
- The capital of France is Paris
Distant Supervision

- Relation Extraction (Mintz et al., 2009)

  - Freebase
  - IsCapital (Paris, France)
  - IsCapital (London, Britain)

  - Paris is the capital and most populous city of France
  - The capital of France is Paris
Dataset Creation

What is Knowledge Base for our task?
Dataset Creation

What is Knowledge Base for our task?
What is Knowledge Base for our task?
Dataset Creation

Attributes we focus on:
- Education
- Job
- Spouse
Attributes we focus on:

- Education
- Job
- Spouse
Dataset Creation

- Education: Positive Examples

Google Plus API
Dataset Creation

- Education: Positive Examples

- Google Plus API → Seeds
  - StudyAt
  - (UserName, SchoolName)
  - (Miranda Cosgrove, USC)
Education: Positive Examples

Google Plus API → Seeds → StudyAt (UserName, SchoolName) (Miranda Cosgrove, USC) → Twitter Search
Dataset Creation

Education: Positive Examples

Google Plus API → Seeds → StudyAt (UserName, SchoolName) (Miranda Cosgrove, USC)

Twitter Search

Results for from:MirandaCosgrove USC

Miranda Cosgrove @MirandaCosgrove · 14 Jan 2013
Took Penelope with me everywhere today. Just got my books at #trojanslovepuppies instagam/p/Ue-vQovsx-
Expand

Miranda Cosgrove @MirandaCosgrove · 10 Dec 2012
Study session at USC! Finals #letsskillit :) instagam/p/TFaOY4vs5c/
Expand
Dataset Creation

- Education: Negative Examples

Google Plus API → Seeds

1. StudyAt
   - (UserName, SchoolName)
2. (Miranda Cosgrove, USC)
**Dataset Creation**

- **Education: Negative Examples**

  - **Google Plus API** → **Seeds** → **Twitter API**
  - **StudyAt**
    - **(UserName, SchoolName)**
  - **(Miranda Cosgrove, USC)**

Jiwei Li, Alan Ritter and Eduard Hovy
Dataset Creation

- Education: Negative Examples

![Diagram showing the process of dataset creation involving Google Plus API, Seeds, Twitter API, and a Profile summary of Miranda Cosgrove](image)

Negative Examples

Jiwei Li, Alan Ritter and Eduard Hovy
Dataset Creation

Education: Data Expansion

Google Plus API \(\rightarrow\) Seeds

StudyAt
(UserName, SchoolName)
(Miranda Cosgrove, USC)

Network Information
Dataset Creation

- Education: Data Expansion

Google Plus API → Seeds

Google Plus API

StudyAt
(UserName, SchoolName)
(Miranda Cosgrove, USC)

Network Information

More Users
Dataset Creation

- Education: Data Expansion

Google Plus API

Seeds

StudyAt
(UserName, SchoolName)
(Miranda Cosgrove, USC)

Network Information

More Users

Jiwei Li, Alan Ritter and Eduard Hovy
Dataset Creation

- Spouse

Freebase API
Dataset Creation

- **Spouse**

```
[Freebase logo]  →  IsSpouse
                  (Tom Cruise, Katie Holmes)
```

Freebase API
Dataset Creation

- Spouse

Freebase API

IsSpouse
(Tom Cruise, Katie Holmes)

Jiwei Li, Alan Ritter and Eduard Hovy
## Dataset Creation

<table>
<thead>
<tr>
<th></th>
<th>Education</th>
<th>Job</th>
<th>Spouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Users</td>
<td>7,208</td>
<td>1,806</td>
<td>1,636</td>
</tr>
<tr>
<td>#Users Connected</td>
<td>6,295</td>
<td>1,407</td>
<td>1,108</td>
</tr>
<tr>
<td>#Edges</td>
<td>11,167</td>
<td>3,565</td>
<td>554</td>
</tr>
<tr>
<td>#Pos Entities</td>
<td>451</td>
<td>380</td>
<td>3121</td>
</tr>
<tr>
<td>#PosTweets</td>
<td>124,801</td>
<td>65,031</td>
<td>135,466</td>
</tr>
<tr>
<td>#Aver Pos Tweets User</td>
<td>17.3</td>
<td>36.6</td>
<td>82.8</td>
</tr>
<tr>
<td>#Neg Entity</td>
<td>6,987,186</td>
<td>4,405,530</td>
<td>8,840,722</td>
</tr>
<tr>
<td>#Neg Tweets</td>
<td>16,150,600</td>
<td>10,687,403</td>
<td>12,872,695</td>
</tr>
</tbody>
</table>

Table 1: Statistics for our Dataset
Algorithm
Potential Function

Given an entity $e$ recognized by Twitter NER (Ritter et al., 2011).

$\Psi(z_i,e)$: Potential function, entity $e$ constitutes the correspondent attribute of user $i$

$$\psi(z_i,e) = \frac{1}{Z} \psi_{Text}(z_i,e) \psi_{Network}(z_i,e)$$
Learning

- **Text-Level Evidence** $\Psi_{Text}(z_{i,e}^k)$
  - Entity-level: number of tokens, capital letter, length
  - Token-level: identity, shape, POS, NER
  - Window-level: tokens, POS
  - Tweet-level: tokens
  - External Sources: list of universities/companies

- **Neighboring Effect**
Learning

- Text-Level Evidence $\Psi_{Text}(z^k_{i,e})$
- Neighboring Effect
  - Education, Job (Homophily)
    \[
    \Psi_{Network}(z_{i,e}) = \prod_{j \in \text{Neigh}(i)} \exp(\lambda I(Z_{j,e} = 1)/N)
    \]
  - Spouse
    \[
    \Psi_{Network}(z_{i,e}) = \exp(\lambda I(Z_{e,\text{user},i} = 1))
    \]
- Max-Ent for training.

Jiwei Li, Alan Ritter and Eduard Hovy

User Profile Extraction
Learning

- Text-Level Evidence $\Psi_{Text}(z^k_{i,e})$
- Neighboring Effect
  - Education, Job (Homophily)
    \[ \Psi_{Network}(z_{i,e}) = \prod_{j \in Neigh(i)} \exp(\lambda I(Z_{j,e} = 1)/N) \]
  - Spouse
    \[ \Psi_{Network}(z_{i,e}) = \exp(\lambda I(Z_{e, user_i} = 1)) \]
- Max-Ent for training.
Learning

- Text-Level Evidence $\Psi_{Text}(z^k_{i,e})$
- Neighboring Effect
  - Education, Job (Homophily)

$$\psi_{Network}(z_{i,e}) = \prod_{j \in \text{Neigh}(i)} \exp(\lambda I(Z_{j,e} = 1)/N)$$

- Spouse

$$\psi_{Network}(z_{i,e}) = \exp(\lambda I(Z_{e,user_i} = 1))$$

- Max-Ent for training.
**Inference**

- **Observed**: Neighboring information is already given (Education, Job).

![Inference Diagram](image)
Inference

- **Observed**: Neighboring Information is already given (Education, Job).

- **Latent**: Neighboring Information is unknown (Joint Inference)
Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**
Latent: Neighboring Information is unknown (Joint Inference)

- Initializing only based on text-level information $\psi_{Text}(z_i, e)$
Latent: Neighboring Information is unknown (Joint Inference)

- Initializing only based on text-level information $\Psi_{Text}(z_i,e)$
- Infer each individual given its neighbors
Latent: Neighboring Information is unknown (Joint Inference)

- Initializing only based on text-level information $\psi_{\text{Text}}(z_i, e)$
- Infer each individual given its neighbors
**Inference**

- **Latent: Neighboring Information is unknown (Joint Inference)**
  - Initializing only based on text-level information \( \Psi_{\text{Text}}(z_i,e) \)
  - Infer each individual given its neighbors
Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**
  - Initializing only based on text-level information $\Psi_{Text}(z_i, e)$
  - Infer each individual given its neighbors
Experiments
Only-Text:
Text-Level Evidence $\Psi_{Text}(z_{i,e})$
Baselines

- Only-Text:
  Text-Level Evidence $\Psi_{Text}(z_i,e)$

- NELL: Bag of words matching in the list of universities or companies borrowed from NELL
Results

![Graphs showing Recall, Precision, and F1 scores for different methods: Observed, Latent, Only-Text, and Nell.](image)

- **Recall**
- **Precision**
- **F1**

Jiwei Li, Alan Ritter and Eduard Hovy

User Profile Extraction
Results

Recall

Precision

F1
Results

![Graphs showing Recall, Precision, and F1 scores for different methods: Observed, Latent, Only-Text, and Nell.](image)
Results

![Graphs showing recall, precision, and F1 scores for different methods: Observed, Latent, Only-Text, and Nell.](image-url)
Conclusion

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion
We present a framework to extract user attributes from Twitter.

We present a large-scale dataset for this task.

We demonstrate the benefit of jointly reasoning about users’ social network structure.
Future Work

Facebook:
Dataset

http://aclweb.org/aclwiki/title=Profile_data
Thank you

Dataset
http://aclweb.org/aclwiki/title=Profile_data

Thank You!

Questions, Suggestions?