Privacy, Online Data, and the JobSeeker

Raquel L. Hill
School of Informatics and Computing
Indiana University, Bloomington, IN

Funded by National Science Foundation: 1537768
Introduction

• While evaluating job applications, recruiters try to determine whether:
  – the information that is provided by a job seeker is accurate
  – it describes a person with sufficient skills

• Prior Research has shown this process to be fraught with bias.
Resume Experiments and Bias

Tweak demographic characteristics

(Levin 2004)

Help Wanted (depending on religious affiliation)

(Acquisti 2015)

Racism in a resume
Job applicants with African American-sounding names got fewer callbacks.

Moss-Racusin 2012)
Possible Solution

• Anonymize Resume
  – Remove all identifying information
• Is this a simple or hard task?
• What makes information identifying?
Uniqueness and Re-Identification

- Uniqueness is required, but is not a sufficient condition for re-identification.
- To re-identify humans in a dataset, uniqueness must be linked with outside knowledge.
The “Re-Identification Problem”

- Ethnicity
- Visit date
- Diagnosis
- Procedure
- Medication
- Total charge
- Hospital Discharge Data

Quasi-Identifier

Name
Address
Date registered
Party affiliation
Date last voted
Voter List

Re-identification of William Weld

5-Digit Zip Code

+ Birthdate

63-87% of USA estimated to be unique

## The Search Log Case (2006)

- **Goal:** Support web search research
- **Customers:** 650k customers, 20 million queries, 3 month period
- **Names:** Replaced with persistent pseudonyms

<table>
<thead>
<tr>
<th>Name</th>
<th>Query</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Doe</td>
<td>Books</td>
<td>1/2/05</td>
<td>16:52</td>
</tr>
<tr>
<td>Bob Smith</td>
<td>Payscale</td>
<td>1/4/05</td>
<td>23:41</td>
</tr>
<tr>
<td>John Doe</td>
<td>Porn</td>
<td>1/8/05</td>
<td>03:15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User</th>
<th>Query</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>8123</td>
<td>Books</td>
<td>1/2/05</td>
<td>16:52</td>
</tr>
<tr>
<td>9010</td>
<td>Payscale</td>
<td>1/4/05</td>
<td>23:41</td>
</tr>
<tr>
<td>8123</td>
<td>Porn</td>
<td>1/8/05</td>
<td>03:15</td>
</tr>
</tbody>
</table>
Barbaro & Zeller. A face exposed for AOL searcher no. 4417749. 

**User 4417749 issued hundreds of searches:**
- Numb fingers
- Dog that urinates on everything
- Hand tremors
- 60 single men
- Last name = "Arnold"
- Landscapers in Lilburn (Georgia)
- Homes sold in shadow lack subdivision Gwinnett county Georgia
- Dry mouth
- Nicotine effect on the body
- Bipolar

*Thelma Arnold & Dudley*
Thelma Arnold & Dudley
Related Work: Re-Identification

• To re-identify humans in a dataset, uniqueness must be linked with outside knowledge.
  – Sweeney (1997): Link gender, birthdate, zipcode with a hospital's discharge records
  – Malin & Sweeney (2001): Link DNA sequences with multiple hospital discharge records
Re-Identification

- Malin (2006): Link genealogical data with online death records and obituaries
- Narayanan and Shmatikov (2008): Link movie ratings and watch dates with attacker's own knowledge, or IMDb forum posts
Social Media and New Identifiers

• There is so much data being self-reported in social media outlets
• Our preliminary studies show:
  – 70% of individuals self-reporting about medical related issues use weak privacy settings
  – 27% use the strongest possible settings
  – 3% use settings that reveal more than name and gender, but less than the maximum amount of information
Social Media Cont.

• (Chaabane 2012) “You are what you like, Information Leakage through User Interests”
  – Age, gender, relationship status, country level location

• Acquisti (2015):
  – Resume study with online data
  – Created social media profile that indicated religious preference
Online Data and Hiring
Concerns

• Jobseekers
  – Try to hide their age
  – Obfuscate career transitions
  – Conceal qualifications
    • Limited training
    • Over qualified
  – Lack tools to express and present their skills

• Employers
  – Seek to identify discrepancies in applications
  – Lack automated tools to evaluate applications
Next Steps

• Move Beyond Resume Experiments

• Build a Framework that:
  – Uses online and resume information to create jobseeker profiles
  – Develop means for identifying discrepancies in information in order to provide better feedback to individuals
  – Enable semantic comparisons between profiles
Our Approach

Create a skills ontology
- Process for creation must be automated
- Manual creation is costly – people
- 100 % NLP is costly (semantic approach)
- We leverage wikipedia structure.
Our Approach

• Develop a skills ontology
  – hierarchy of job skills
  – relationships between the specified skills
• The proposed framework uses the web as its language corpus
• Data mining approach to the problem
Challenges

• Hidden skills problem:
  – Concerns representing, identifying and measuring skills that haven't been explicitly mentioned in the job description or resume
    • For example a job description may state that the applicant should know PHP, and an applicant's resume lists Symfony.

• Skills Resolution problem:
  – Involves taking text from online social media and online blogs and mapping that text to actual skill terms
Meeting the Challenge

• Any automatic software aiming to help the hiring process and address the mentioned challenges should benefit from skills taxonomy

• A comprehensive skills taxonomy is needed in order to:
  – represent the skill set of job applicants
  – help the employers to describe job description in well-defined way
  – measure the distance between their skills set and the required skills of the ideal applicant
Motivating Example

• Let’s see how a simple flat dataset of skills (related to IT) can help automatic tools
  – For Skill resolution problem
• Used Stackoverflow API to retrieve posts of users
• Aim is to create skill clouds for some active users in Stackoverflow
  – Derive skills from demonstrated knowledge
1st Experiment

• Keyword Cloud using posts authored by user
2nd Experiment

- Tag Cloud using frequent tags in posts authored by user
  - Just using top frequent tags
  - No filtering
3rd Experiment

- Tag Cloud using frequent associated tags in posts with Skill dataset
  - Filtering is considered using simple skill dataset
How to Create Skills Taxonomy

• Automatic tool to develop a skills taxonomy
  – capable of automatically bootstrapping the taxonomy
• Bootstrapping:
  – a process for learning relationship rules that alternates between learning rules or rule accuracy from sets of instances of included entities and finding instances using sets of rules
• We will use a data mining approach to this problem
• Input
  – minimal skills taxonomy
  – Web
• Output:
  – More comprehensive taxonomy
Bootstrapping Taxonomy

• We will benefit from 3 approaches:
  – Word2Vec
    • Learn language model from a text corpus and find similar words based on learned model
    • No semantic relationship between words
  – Wikipedia Structure
    • Can learn some semantic relationship
    • Article titles are not always good enough for skills but relevant
  – Search Engine Result mining
    • Broader source of knowledge but with more noise
    • Can be used for learning semantic relationship but with lots of noise
    • Can benefit from searching in job descriptions
Exploiting Wikipedia Structure

- Wikipedia as a source of knowledge
  - large corpus containing millions of articles about named entities
  - Lots of knowledge
  - Embedded structure is not enough to be used directly for ontologies
- Relationship between articles can be defined in many ways
- One way is to use Wikipedia "category":
  - Relationship between articles
  - These relationships may not be the “Is-a relation”, which introduces noise into the relationship structure
Creating Wikipedia Graph

- Using the Wikipedia API we started scraping from one article with the title 'Python (programming language)'.
- Using categories relationships
  - Scraped parent categories
  - Subcategories of each category in the path
  - Stopping condition: 2 hops away from source
- The resulted undirected graph had 1359 nodes and 1581 edges
- Manual annotation of the article titles indicates:
  - 80 percent of the nodes are non-relevant to skills and 20 percent have some information related to skills
Skills Graph
Graph Mining

- Simple feature engineering to create a structured table for classification algorithm

<table>
<thead>
<tr>
<th>Feature</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree</td>
<td>number of adjacent nodes</td>
</tr>
<tr>
<td>AVG degree of neighbors’</td>
<td>average degree of adjacent nodes</td>
</tr>
<tr>
<td>exist in k-core</td>
<td>whether node exist in 2-core subgraph</td>
</tr>
<tr>
<td>No of neighbours as skills</td>
<td>number of adjacent nodes within the path equal to 2</td>
</tr>
<tr>
<td>Path to skill</td>
<td>minimum path to skill node</td>
</tr>
<tr>
<td>Cluster no</td>
<td>the id of the cluster that node is assigned</td>
</tr>
<tr>
<td>Cosine similarity</td>
<td>maximum cosine similarity value with skills in dataset</td>
</tr>
<tr>
<td>No of similar skills</td>
<td>number of skills that the cosine similarity value is above threshold</td>
</tr>
<tr>
<td>target column</td>
<td>article is related to IT skills</td>
</tr>
</tbody>
</table>
Performance of Classifiers

- We randomly chose 80% of dataset as the training set and 20% as the test set
- We used cross validation with 10 folds
- Maximum likelihood estimator is considered as baseline
  - Meaning prediction of mode (not related to skill) for all records

<table>
<thead>
<tr>
<th>algorithm</th>
<th>Acc.</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.80</td>
<td>80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.81</td>
<td>0.78</td>
<td>0.37</td>
<td>0.53</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.98</td>
<td>0.98</td>
<td>0.94</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Future Work

• Experiments related to search engine results mining

• Test Wikipedia graph with larger graph

• Test Meta Classifier and analyze final results

• Run skill cloud generation with the final skills taxonomy
Thanks

• National Science Foundation

• Collaborators: Mohsen Sayyadi, Ilana Gershon