Introduction to Entity Search

Web Science 2010/2011
Gianluca Demartini

demartini@L3S.de
Topics

• 1. Information Integration
• 2. Web Information Retrieval
• 3. Entity Search
• 4. Web Usage
• 5. Collaborative Web
• 6. Web Archiving
• 7. Medical Social Web
Outline

• From documents to entities
• Different Entity Search tasks
  – Entity Identification
    • Okkam
  – Expert Finding
    • In a company
  – Entity Ranking
    • In Wikipedia
    • On the Web
• Selected Papers
Entity Search

• Lecture 1: Entities

• Lecture 2: Search
From Documents to Entities

- Document Search
From Documents to Entities

• Entity Search

<table>
<thead>
<tr>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ent1</td>
</tr>
<tr>
<td>2. Ent2</td>
</tr>
<tr>
<td>3. Ent3</td>
</tr>
</tbody>
</table>
A taxonomy of Entity Search tasks

- **Document Search**
- **Question Answering**
- **Entity Retrieval**
  - **Navigational**
  - **Informational**
- **Related Entity Search**
  - Query: «Boeing 747»
    - Result: Lufthansa, British Airways, Korean Air, Asiana Airlines
- **Entity List Completion**
  - Query: «Nobel Prize Winners»
    - Albert Einstein, Renato Dulbecco
  - Query: Maria Skłodowska-Curie, Dalai Lama, Barack Obama
- **People Search**
- **Movie Search**
- **Country Search**
  - Query: «Countries where I can pay in Euro»
    - Result: Germany, Italy, Spain, France
- **Expert Finding**
  - Query: «C++ compilers»
    - Bjarne Stroustrup, John Doe
Entity Identification

An open and global service which can be used within existing applications to support the creators/editors of semantic web content to (re)use the same globally unique URI for referring to the same entity in a systematic way.

- Okkam

- Sig.ma
  - Building Entity Profiles
Outline

• From documents to entities
• Different Entity Search tasks
  – Entity Identification
    • Okkam
  – Expert Finding
    • In a company
  – Entity Ranking
    • In Wikipedia
    • On the Web
• Selected Papers
Expert Finding

• Scenario:
  – Executives need to create a team for a new project: find staff with the right expertise
  – Someone needs to solve a problem

• Goal:
  – Use the digital content available in the enterprise
  – Create a ranking of people who are experts in the given topic
Evidence of Expertise

- Email or bulletin board messages
- Corporate communications
- Shared folders in file system
- Resumes and homepages
- Employee database
- Email flow
- Bibliographic information
- Software library usage
- Search and publication history
- Project time charges

See also bibliography on TREC-ENT wiki:
Two Basic Approaches

Who should I ask about the copyright forms?

- Document-based: rank docs, extract experts

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copyright forms

1. Lori
   | Lori
2. Lori
   | Ellen
3. Ian
   | Lori
4. Lori
5. Ellen
   | Lori
Document-based Expert Finding

• Find and score documents about the topic
  – Title about topic
  – Abstract about topic

• Aggregate scores for each distinct author
Two Basic Approaches
Who should I ask about the copyright forms?

• Document-based: rank docs, extract experts

  Copyright forms

  1. Lori
     2. Lori
     3. Lori
     4. Lori
     5. Lori

• Candidate-based: rank candidate profiles

  Copyright forms

  Lori
  Ellen
  Ian
Additional Techniques
Research Systems

• Combine the two basic approaches
• Estimate the quality of the evidence
• Use of collection/structural knowledge
  – Treat emails different from documents
  – Treat email’s subject/sender/receiver different from body
  – Locate homepages
• Use social network extracted from co-authorship or email lists

See also TREC proceedings 2005-2007
Key Requirements

• **Identify** experts via self-nomination and/or automated analysis of expert communications, publications, and activities

• **Classify** the type and level of expertise of individuals and communities

• **Validate** the breadth and depth of expertise of an individual

• **Recommend** experts, including the ability to rank order experts on multiple dimensions including skills, experience, certification and reputation
Evaluating Expert Finding Systems

• TREC Enterprise track 2005-2008
  – http://www.ins.cwi.nl/projects/trec-ent/

• Standard test collection using
  – W3C website
  – CSIRO website

• Queries and manual relevance judgements

• Evaluation measures to compare systems
Outline

• From documents to entities
• Different Entity Search tasks
  – Entity Identification
    • Okkam
  – Expert Finding
    • In a company
  – Entity Ranking
    • In Wikipedia
    • On the Web
• Selected Papers
Ranking...

- People
- Actors
- ... Car companies
  
  [i.e., insert your fav entity type here]

Entity Ranking!!!
Examples of *Entities* in Wikipedia

- Art museums and galleries
- Countries
- Famous people
- Monarchs of the British Isles
- Artists
- Magicians
Example Entity Ranking Scenarios

• Impressionist art museums in Holland
• Countries with the Euro currency
• German car manufacturers
• Artists related to Pablo Picasso
• Countries involved in WWI
• Actors who played Hamlet
• English monarchs who married French women

Many examples on
http://www.ins.cwi.nl/projects/inex-xer/topics/
Entity Ranking

- Topical query $Q$
- Entity (result) type $T_x$
- A list of entity instances $X_s$

- An entity is represented by its Wikipedia page
- Systems employ categories, structure, links
Tasks

• Entity Ranking (ER)
  – Given Q and T, provide Xs

• List Completion (LC)
  – Given Q and Xs[1..m]
  – Return Xs[m+1..N]
The user wants the dinghy classes that are or have been olympic classes, such as Europe and 470.

The expected answers are the olympic dinghy classes, both historic and current. Examples include Europe and 470.
Formal Model for Entity Ranking

Indexing

- Entities
- Data Sources

“Alexandre Pato”
ID: ap12dH5a
(born in; 1989)
(playing with; acm15hDJ)
Formal Model for Entity Ranking

- Searching
  - Users' Information Need
  - Entity Ranking System

1. User Query
   - Keywords or Natural Language Text

2. Processed Query
   - List of Attribute-Value pairs

3. Rank Entities
   - Use scoring function

4. Entities

\[ E = \{e^1, \ldots, e^i\} \]

\[ E = \{id(e^i), \phi(q, d(e^i))\} \]
Wikipedia

• Encyclopedia
  – multilingual, Web-based, free-content, openly-editable: errors are promptly corrected

• Articles:
  – balanced, neutral, and encyclopedic, containing notable verifiable knowledge

• Categories / sub-categories

• Links, anchor text (Germany -> Albert Einstein)
Approaches to ES in Wikipedia

• Exploit and refine the category structure
  – Wordnet to find entity types (e.g., a professor is a person)
• Extend the query
  – Synonyms and related words (Wordnet synsets)
• Exploit the link structure
  – Links in Wikipedia are usually entities
  – Search Keywords also in anchor text of outLinks
YAGO

- Suchanek et al. 2007
- Highly accurate ontology (>95%)
- Extracted from Wikipedia + WordNet
- Provides semantic concepts describing Wikipedia entities
Category Based Search

• Query expansion by modifying category information
  – Subcategories
    • Extracted from Wikipedia
  – “Children” Categories
    • Filtered using the YAGO subClassOf relation
  – “Sibling” Categories
    • Extracted from Wikipedia
    • Having with the same YAGO type
“Children” Categories

- Situation Comedy
- YAGO `subClassOf`
- BBC Television Sitcoms
- Latino Sitcoms
- Sitcoms in Canada
- Fictional Character

Wikipedia Subcategories

- Sitcoms
- Wikipedia Category
“Sibling” Categories

Situation Comedy

YAGO subClassOf

BBC Television Sitcoms
Latino Sitcoms
Sitcoms in Canada
...

YAGO subClassOf

Sitcoms
Wikipedia Category
Evaluating ES in Wikipedia

• INEX Entity (XER) track 2007-2009

• Standard test collection using
  – Wikipedia dump from 2006
  – Wikipedia dump from 2009 + extracted entities and types from Wordnet

• Queries and manual relevance judgements

• Evaluation measures to compare sytems
Ranking Entities on the Web

• TREC Entity Track 2009-2010
  – 50M web pages (including Wikipedia)
  – Find related entities (return homepages)

<query>
  <num>7</num>
  <entity_name>Boeing 747</entity_name>
  <entity_URL>clueweb09-en0005-75-02292</entity_URL>
  <target_entity>organization</target_entity>
  <narrative>Airlines that currently use Boeing 747 planes.</narrative>
</query>
Ranking Entities on the Web

• Approaches
  – Use Wikipedia (and infoboxes) as background info
  – Extract entities from tables and lists
  – Find the homepage given the entity name (see ENS)
    • Barack Obama -> www.barackobama.com

• In 2010: 1 billion web pages
Time-Aware Entity Retrieval

• In some cases the time dimension is available
  – News collections
  – Blog postings

• An Entity Search system can exploit the past to find relevant entities
Time-Aware Entity Retrieval

User → news query → H → summary ranking entities → e1, e2, e3, ...

Charles Schulz Dies

Important Entities:
- Charles Schulz
- Congressional_Gold_Medal
- Santa_Rosa
- Peanuts

AP Online
02-15-2000
House Honors 'Peanuts' Creator

WASHINGTON (AP) -- "Peanuts" creator Charles Schulz was remembered today as a genius who touched the lives of millions of Americans as the House adopted a resolution to award him a Congressional Gold Medal.

The 77-year-old cartoonist died in his sleep Saturday at his Santa Rosa, Calif., home, a day before Schulz's last strip featuring Snoopy and the gang was published. He had announced in November he would retire after being diagnosed with colon cancer.

"On Saturday night, millions of Americans lost their security blanket," said Rep. Lynn Woolsey, D-Calif. "Life won't be the same without Charles …

ES Commercial systems

• Entity Ranking
  – http://www.google.com/squared/search?q=german+beers

• List Completion
  – http://labs.google.com/sets?hl=en&q1=ferrari&q2=mbw&q3=mercedes&q4=&q5=&btn=Small+Set+(15+items+or+fewer)

• Related entities
  – http://correlator.sandbox.yahoo.net/index.php/concepts/beer
Outline

• From documents to entities
• Different Entity Search tasks
  – Entity Identification
    • Okkam
  – Expert Finding
    • In a company
  – Entity Ranking
    • In Wikipedia
    • On the Web
• Selected Papers
Selected Papers


Formal models for expert finding in enterprise corpora

- Expert Finding task
- Compares the two approaches:
  - Model experts based on the associated docs
  - Locate relevant documents, then finds the experts
- Defines probabilistic models
Why finding entities in Wikipedia is difficult, sometimes

• ER and LC in Wikipedia
• Uses WordNet (see Introduction to Information Integration)
  – Refining the category structure
• Rewrites the query using Natural Language Processing techniques

• Demo: http://serwi.L3S.uni-hannover.de
Entity Ranking using Wikipedia as a Pivot

• Related Entities task
• First finds relevant web pages
• Then finds relevant entities using Wikipedia “external links” and types
• Can deal with most of (but not all) the queries
Ranking Related Entities: Components and Analyses

- Related Entities task

- High recall (can find most of the relevant entities)
- Problems with ranking (entities of the wrong type are returned)
¿Questions?

• demartini@L3S.de
• Room 240, Appelstr. 4