

# Researcher affiliation extraction from homepages

**I. Nagy, R. Farkas, M. Jelasity**  
**University of Szeged, Hungary**





# Scientific social information

- **Research interest**
- **Education**
- **Previous and current affiliations, projects**
- **Professional memberships**
- **Teaching activities**
- **Students, supervisors**
- **Personal (nationality, age)**



# Source of social information

- **Social sites (e.g. LinkedIn)**
  - structured, lots of information
  - coverage?
- **Citation databases**
  - limited information (coauthors, affiliations, citations)
- **Homepages**
  - thought to be important by the researcher himself
  - almost every researcher has (a) homepage
  - unstructured



# Web Content Mining

- **Early systems ('99-2000): expert rules**
- **Seed-driven systems**
  - Input: seed pairs of target information
  - Extract patterns from unlabeled text (e.g. Web)
  - Exploits redundancy (celebs)
  - High precision
- **Researcher homepage**
  - Long tail, high recall required

# Case study: affiliation information

- [*affiliation; position; start date; end date*]
- **Frequently given**
- **Experiences can be generalised**
- **useful:**
  - **Collegial relationships (whether they worked with the same group at the same time)**
  - **Do American or European researchers change their workplace more often?**





# Architecture

- **Locating the homepage of the researcher**
  - name disambiguation
- **Locating the relevant parts of the site**
  - pages (focused crawling), parts
- **Extracting information tuples**
  - Weakly supervised setting
- **Normalisation**
  - For every source of information



# Manually tagged corpus

- **455 sites, 5282 pages for 89 researchers**
- **three-level deep annotation hierarchy with 44 classes**
- **manual annotation in the original HTML format (WYSWYG) with hyperlinks**
- **low inter-annotation agreement**
- **focus on affiliation**

# Sample

## Curriculum Vitae

Born on June 12, 1968 in Karlsruhe, Germany, Professor Dr. rer. nat. Alexander Keller studied computer science at the University of Kaiserslautern from 1988 to 1993. He then joined the Numerical Algorithms Group at the same university. Under the supervision of Prof. Dr. S. Heinrich he pursued his Phd-studies and defended his thesis on Friday, the 13th of June, 1997. In 1998 Alexander Keller was appointed scientific advisor of **mental images**. Among four calls in 2003, he chose to become a professor for computer graphics at the University of Ulm in Germany. If time permits, Dr. Funk still joins the gigs of the [Sound Express Big Band](#), where he has been a member for over 20 years.

Search uni-ulm.de for

[Google](#)

## Research

### Main research interests include

- computer graphics (interactive production quality rendering),
- Monte Carlo and quasi-Monte Carlo methods,
- wavelets and the lifting scheme,
- computer vision,
- high performance computing, scientific computing,
- high performance ray tracing, and the
- catalytic conversion of algorithms,



# Textual information

- 47% textual, 24% itemised, 29% hybrid
- **Structured: wrapper induction**
- **Textual paragraph: longer than 40 characters and contains at least one verb**

# researchers	59
# pages	103
# paragraph	151
# sentences	181
# affiliation	374
# position_type	326
# year	212



# Relevant parts

- **Every researcher has (a) homepage**
- **Every homepage can be found in the top10 Google response (query=name)**
- **„CV site” always in depth 1**
- **Textual paragraphs contain cluewords**
  - **class conditional prob. based 1-DNF**
  - **filtering 70k irrelevant paragraphs**

# Slot detection

- **It is not a NER**
  - just affiliation related entities
  - surface features are insufficient
- **Standard procedure (CRF)**
  - with domain specific lists as extra feature
  - domain specific segmentation
  - 70% phrase level F-measure, one-researcher-leave-out  
(37% by lists/regexp)



# Subject detection

- **Sometimes information about supervisors, colleagues**
- **Hypothesis: paragraphs are „homogeneous”**
- **Two procedures**
  - **NER for person names (trained on CoNLL)**
  - **personal pronouns**
  - **~70% accuracy on gold standard and on predicted too**





# Collecting information tuples

- *affiliation* is the head
- Heuristic: assign each *year* and *position\_type* to the nearest *affiliation*
- ~90% accuracy using the gold-standard labels
- ~70% accuracy using the labels predicted by the system (FPs count as misclassified)

# Problematic issues

- „*I am a Ph.D. Student working under the supervision of Prof. NAME*”
- „*Hewlett-Packard Labs **in Palo Alto***”
- „*Ph.D. from MIT **in Physics***”
- „*Department of Computer Science, [Waterloo University]*<sub>BASELINE</sub>”
- „*I **lead** the Distributed Systems Group*”
- **In-domain name detection**
- **Enumeration detection is important (syntactic parsers?)**



# Conclusions

- **Information from homepages of researchers**
- **Special nature of the tasks:**
  - long tail
  - small labeled corpus
  - lack of domain-specific parsers
- **Several well defined subtasks**
- **Basic solutions for each subtask**



**Thank you!**

[www.inf.u-szeged.hu/rgai/homepagecorpus](http://www.inf.u-szeged.hu/rgai/homepagecorpus)  
[rfarkas@inf.u-szeged.hu](mailto:rfarkas@inf.u-szeged.hu)