Tool support for technology scouting using online sources

Elena Tsiporkova and **Tom Tourwé**

Technology scouting …

… is understood as an organised approach for identifying technological needs, gaps and opportunities, and then finding solutions outside the official borders of the enterprise

Technology scouting …

- … is very often applied when:
  - a technical problem needs to be solved quickly due to some change in the competitive landscape;
  - an organisation is looking for opportunities to move into a new market with limited involvement of internal resources;
  - or specific new skills need to be acquired without increasing internal resource overhead.

- A technology scout needs to utilize an extensive and varied network of contacts and resources, and stay on top of emerging trends and technologies.

An example of application-specific expert finding

- A pharmaceutical company needs to make important decisions related to the planning of large-scale clinical trials world-wide
- Optimal planning is essential since
  - enormous investment is associated with clinical trials
  - the outcome can have crucial impact on the future and profitability of the whole enterprise,
- … and requires
  - thorough knowledge of the different clinical researchers world-wide active in the disease targeted by the planned clinical trials:
    - current research results in the field
    - present affiliation & size of the lab
    - number of patients with the disease in question treated per year
    - …

Going beyond recommendation and personalisation imposes **more strict requirements**

- **Very high (if not complete) coverage** over the domain should be attained
- **High accuracy/reliability** of the data needs to be guaranteed
- The data needs to be **up-to-date** at all times

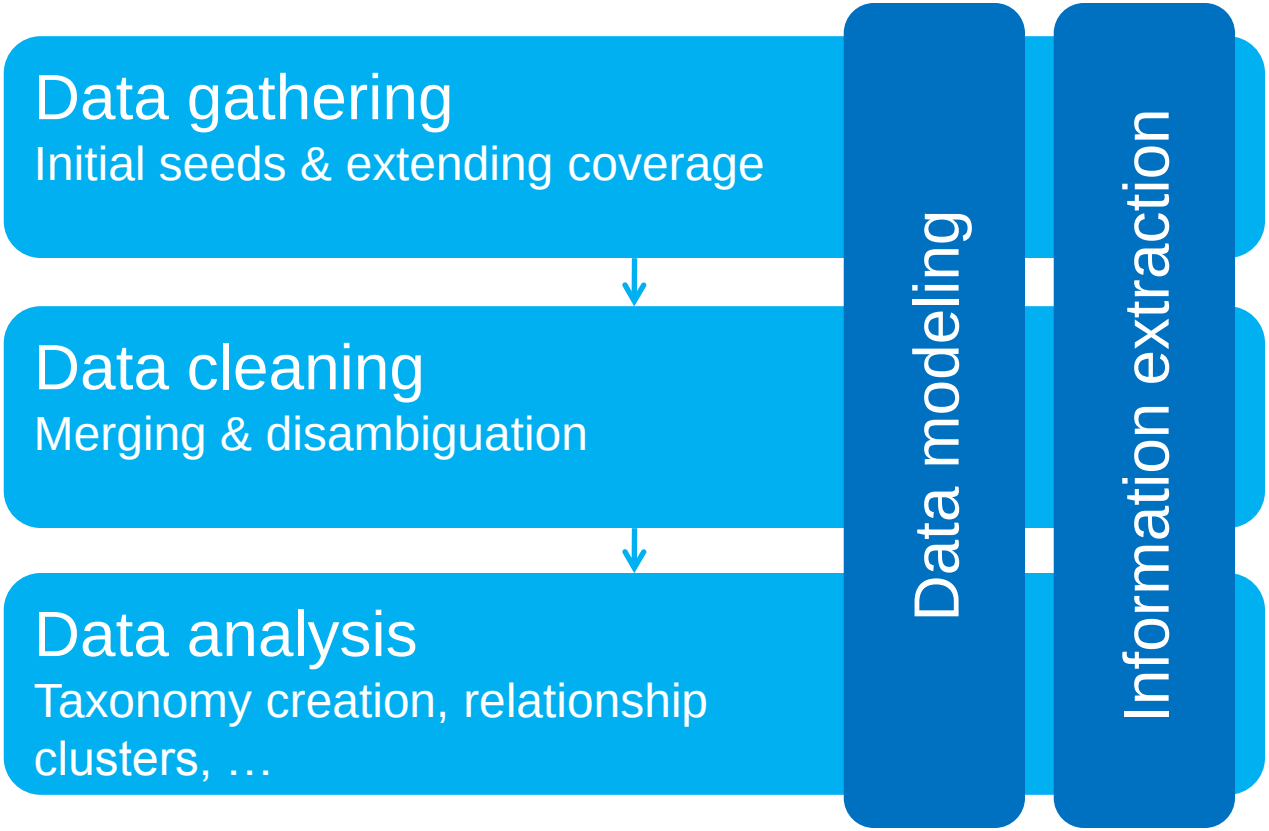
These requirements induce specific **research challenges**

- **High level of coverage** requires information extraction from multiple heterogeneous data sources, structured (LinkedIn, Twitter), semi-structured (DBLP, ACM DL) and unstructured (web pages)

- **High accuracy & reliability** requires
  - to develop and apply advanced disambiguation techniques
  - to qualify the different sources in terms of reliability and trustworthiness: e.g. distinguish between doubtful and reputable sources

- **Keeping the data up-to-date** requires a flexible data model that can deal with partial data and allows continuous updates

A bottom-up approach to building an expert-finding repository



High coverage through **incremental extension of targeted initial seeds**

- Identify online sources to mine serving as seeds for incremental growth of the repository, *targeted to the application domain in question*
- Consider additional sources using the extracted information as a seed
  - *Search for authors and co-authors*
  - *Identify additional published material*

LinkedIn Google scholar beta Microsoft Academic Search Beta dblp.uni-trier.de

Actors in the field

Technology-related publications

Research activities

Career evolution

High accuracy & reliability through **merging & disambiguation**

- The collected data represents partial information about authors, their publications, affiliations, co-authors, …
- This information is often inconsistent and conflicting
  - "T. Tourwé" vs. "Tom Tourwé" vs. "Tom Tourwe" vs. "Tourwé, T." vs. "E. Tourwe" vs. "Tourwé, D"
  - tom.tourwe@sirris.be vs. tom.tourwe@vub.ac.be vs. "tom.tourwe@prog.vub.ac.be" vs. tom.tourwe@cwi.nl
  - "Vrije Universiteit Brussel" vs. "V.U.B" vs. "Brussels Free University" vs. "Free University of Brussels"
  - …
- Merging & disambiguation are required to guarantee that an expert profile and associated publications refer to a unique author

Continuous merging & disambiguation require **flexible data modeling**

- The stream of information is infinite & continuous
  → new information becomes avaiable and is gathered constantly

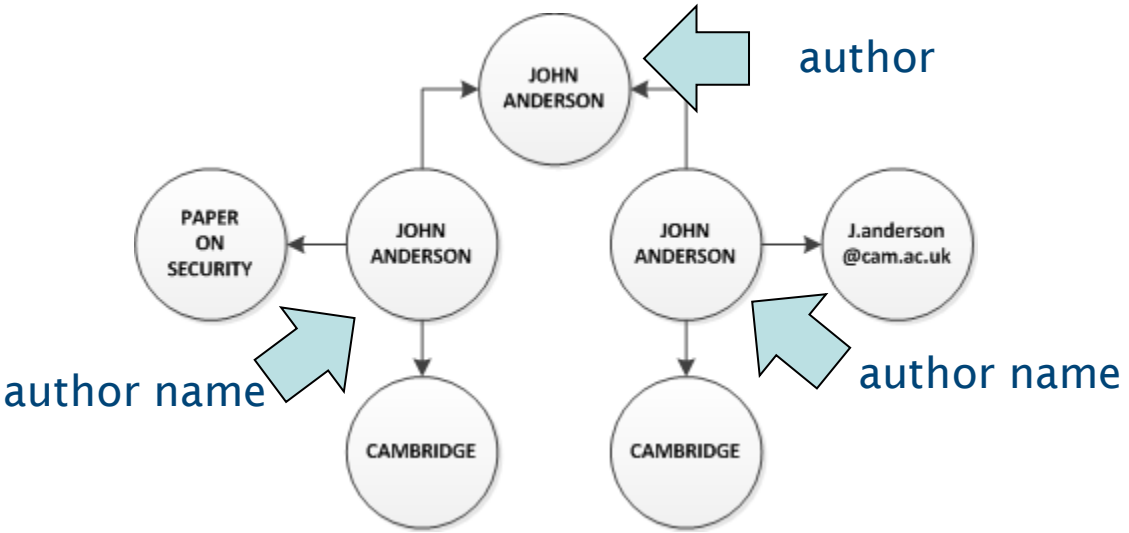
- Information should be considered partial at any moment in time
  → different pieces of information are gathered from different sources and added at different moments in time

- No decision can be permanent
  → decisions made based on partial information might need to be revoked

Graphs as a flexible data model

- Extracted information is represented as an “instance”, a collection of nodes and edges that describe (partial) information about an author

- Constructing a complete author profile amounts to finding an optimal partitioning (clustering) of instances resulting in each instance-group (cluster) representing a unique author

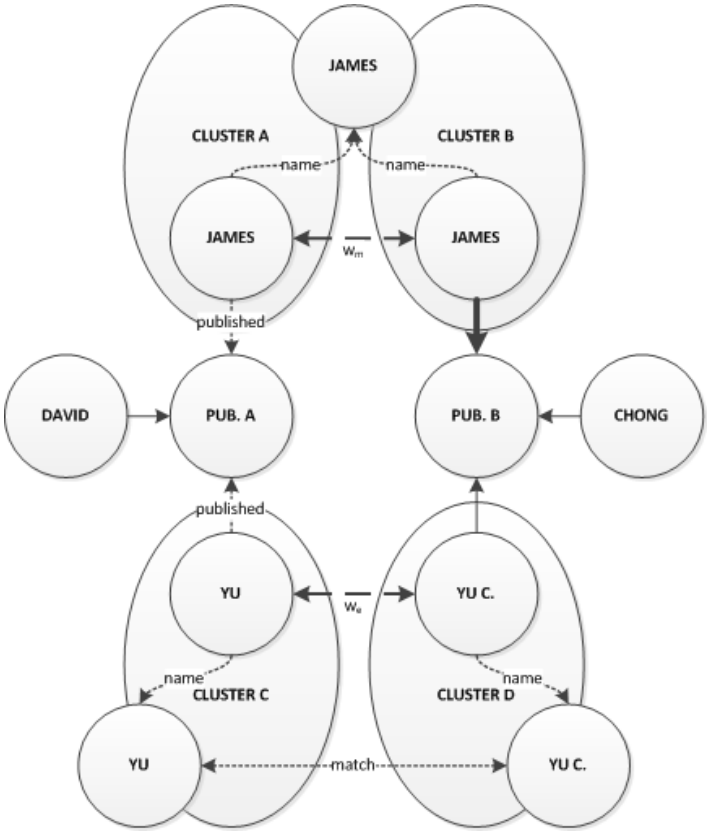


Continuous incremental graph (re)clustering

- Similarity edges are added between name, e-mail address and affiliation nodes
- A domain-independent dynamic minimum-cut tree algorithm computes clusters based on these similarity edges
  - Builds only part of the minimum-cut tree as necessary
    - The number of authors impacted by new data entry is limited
    - The tree is computed over subset of nodes, which affects limited number of clusters
  - Guarantees efficiency while maintaining an identical cluster quality as the static version of the algorithm
- Domain-dependent rules propagate similarities when clustering occurs
  - Community, e-mail & affiliation rules

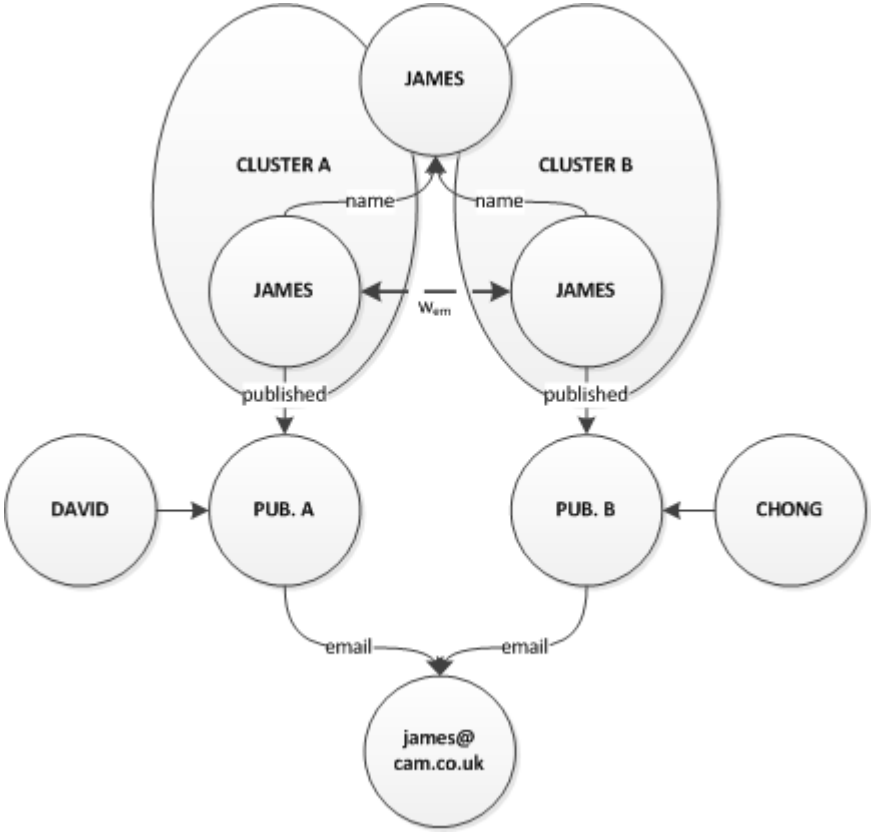
Community rule

- Exploiting the fact that authors often work together with the same co-author



E-mail rule

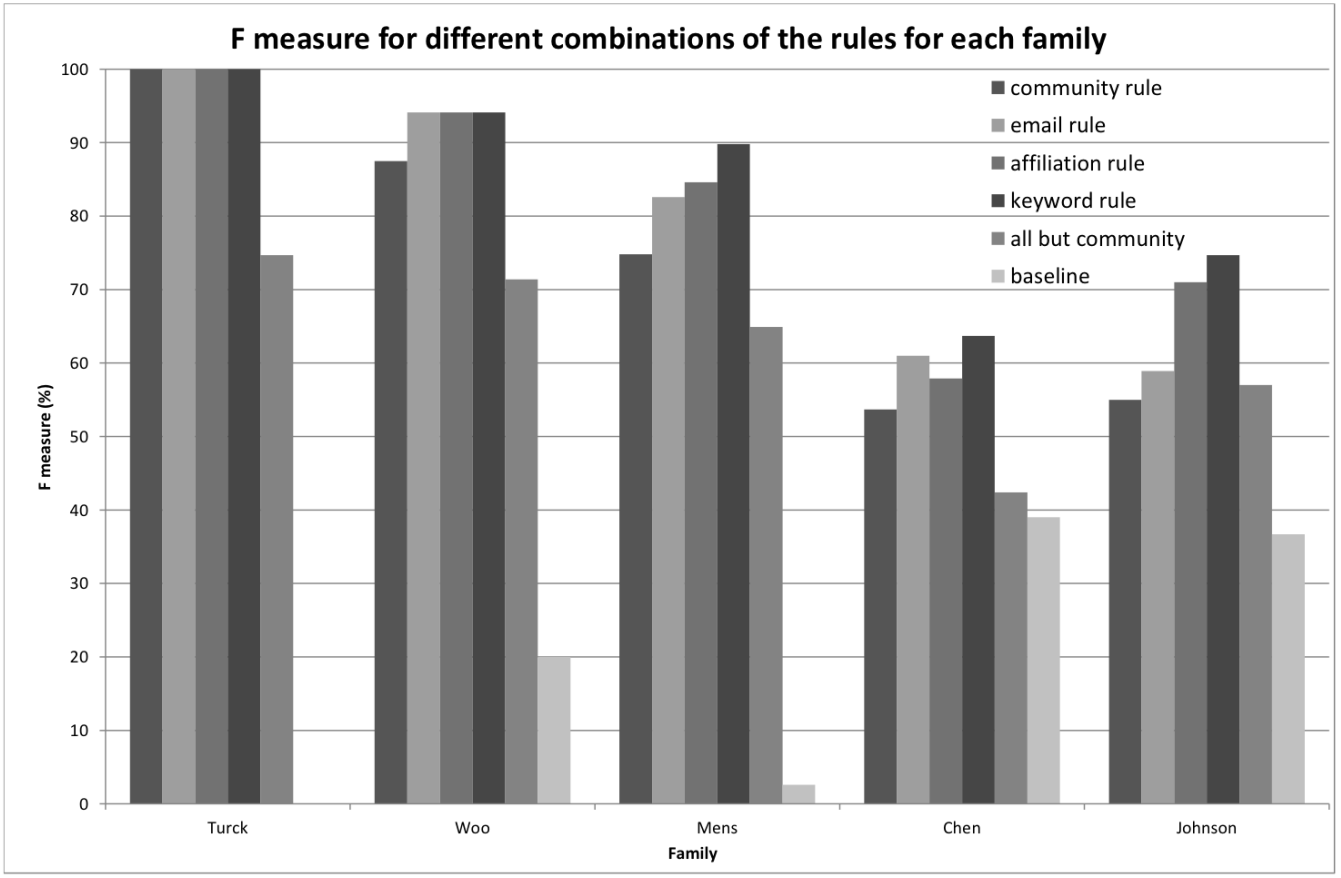
- Authors with the same e-mail address, are most likely the same person



Initial Framework Evaluation

- Manually constructed ground truth testset based on 5 base names and over 1000 publications, extracted from DBLP
  - Turck, Mens, Chen, Woo & Johnson
- Calculcate precision & recall and derived F-measure to compare manually constructed clusters with clusters computed by the algorithm & the different rules

Results


F measure for different combinations of the rules for each family
community rule
email rule
affiliation rule
keyword rule
all but community
baseline
F measure (%)
100
90
80
70
60
50
40
30
20
10
0
Turck
Woo
Mens
Chen
Johnson
Family

Conclusion

- Requirements & research challenges for technology scouting based on an expert-finding repository
- An initial prototype that
  - continuously gathers data based on initial seeds
  - uses a flexible data model
  - incrementally clusters authors based on a domain-independent algorithm and a set of domain-dependent rules
- An initial experiment that evaluates the proposed approach