31. August 2011

www.know-center.at

Model Selection Strategies for Author Disambiguation

TIR'11

Roman Kern, Mario Zechner, Michael Granitzer

Outline

- Problem Statement
- Disambiguation Workflow
- Clustering & Model Selection
- Data Set & Experiments
- Conclusion

Problem Statement
Definition

Given a set of scientific publications/citations, our aim is to **identify distinct authors and their respective publications** within the set

Problem Statement
Examples

Where does ambiguity come from?

- Two distinct authors share the same name
- A single author is referred to by different orthographic variations of her name
- An author has changed her name due to marriage or other causes

3. ☑ Keep +Self -Self 0 Self Citations +Auth -Auth 1 Authors Refine This Author List

[PDF] Hierarchical text classification using methods from machine learning
M **Granitzer** - Master's Thesis, Graz University of Technology, 2003
... Methods from Machine Learning Michael **Granitzer** Page 2. Hierarchical Text Classification using Methods from Machine ... submitted by Michael **Granitzer** Institute of Theoretical Computer Science (IGI), Graz University of Technology A-8010 Graz, Austria 27th October 2003 ...
Cited by 24 - Related articles - View as HTML - Austrian Union Catalog - All 10 versions - Import into BibTeX

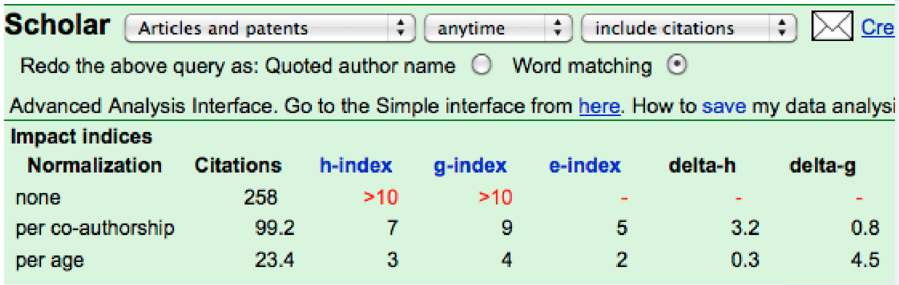
4. ☑ Keep +Self -Self 0 Self Citations +Auth -Auth 3 Authors Refine This Author List

Apical and basolateral conductance in cultured A6 cells
M **Granitzer**, T Leal, W Nagel... - Pflügers Archiv European ..., 1991
1 D6partement de Physiologic, Universit6 Catholique de Louvain, Av. Hippocrate 55, B-1200 Bruxelles, Belgium 2 Physiologisches Institut der Universit/it M/inchen, Pettenkoferstrasse 12, W-8000 Munich 2, Federal Republic of Germany
Cited by 25 - Related articles - All 4 versions - Import into BibTeX

www.know-center.at

Application Scenarios

- Citation and Impact Analysis



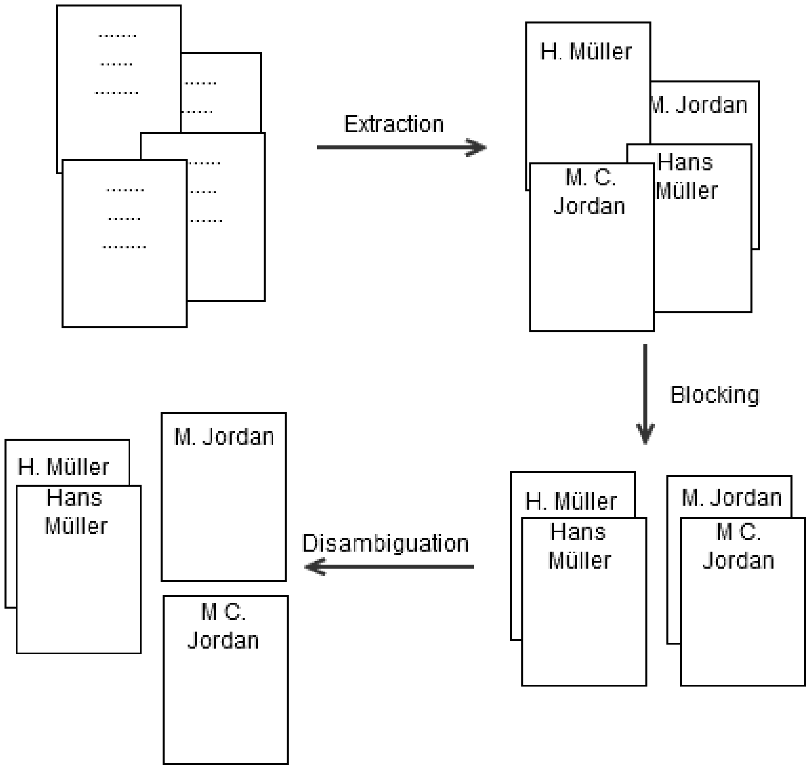
- Creating author profiles in social research networks like Mendeley

- Recommendation engine for research papers
- Facetted Search

A Disambiguation Framework
Overview

Many systems presented in the literature for author disambiguation share the same workflow

1. Extract author name occurances from publications/citations
2. Block author names and the publications they occur in
3. Disambiguate authors within each block



A Disambiguation Framework
1. Author Name Extraction

Extraction of author names

- **Rule-based** extraction based on known publication layouts
- Machine learning techniques for **sequence tagging** (HMMs, CRFs, SVMs)
- **Achievable Performance** (not part of this paper)
  - 0.8-0.9 Precision
  - 0.5-0.8 Recall

➔ The result of this stage is **a set of author names for each publication/citation**

A Disambiguation Framework
2. Blocking

- Blocking is the process of **grouping sufficiently similar author names and the publications/citations** associated with them
- Blocking is performed **for performance** and tractability reasons
- **Similarity measures** for author names
  - Phonetic hashing via Soundex or Metaphone
  - String hashing methods
- **Focus on Recall** – a block must contain all possible unique authors and their publications

A Disambiguation Framework
3. Disambiguation

- Disambiguation is most often achieved via **clustering**
  - For every block
  - Consider all pairs of author names represented as strings and their occurrence in publications
  - Cluster all pairs to groups with unique authors, so that
    - all pairs in a group represent the same unique author
    - All publications of one authors are contained in one group
- Core decisions to apply clustering
  - **Features** to represent pairs
  - **Similarity Measures** between pairs
  - **Model Selection** Method, i.e. guessing the number of authors in one block
  - **Clustering algorithm**

Clustering Properties
Features & Similarity Measure

- Features
  - Noun, adjective and adverbs of publications plain text and information obtained from search engines
  - Keywords extracted from a publiications plain text (TextRank)
  - Tokenized title text
  - Tokenized author names
- Cosine as similarity measure

Clustering Properties
Model Selection – I

- Guess the number of unique authors in one block
  - Large variance, correct number can range from 1 to 100 (and more)
  - Hard problem, often neglected by related work

- Standard methods exists (e.g. density based, stability based)
  - Preliminary test showed very low accuracy

  ➔ **Development of a task specific model selection strategies (main contribution)**

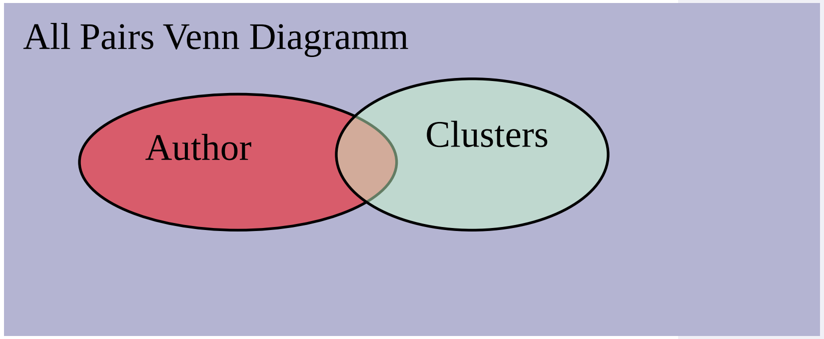
Clustering Properties
Model Selection - II

- Approach/Assumption
  - Clustering groups textually similar publications of a block
  - Use different feature kind for model selection: Co-authorship
  - The more co-authors overlapp in a cluster and the less they are spread between cluster, the better
- Use conditional probabilities as measures therefore

$$\overline{P}_{ac} = \frac{1}{|pairs|} \sum_{\forall pairs} 1 - P(author \mid cluster)$$

$$\overline{P}_{ca} = \frac{1}{|pairs|} \sum_{\forall pairs} 1 - P(cluster \mid author)$$

$$F(C) = (\overline{P}_{ac} + \overline{P}_{ca}) / 2$$



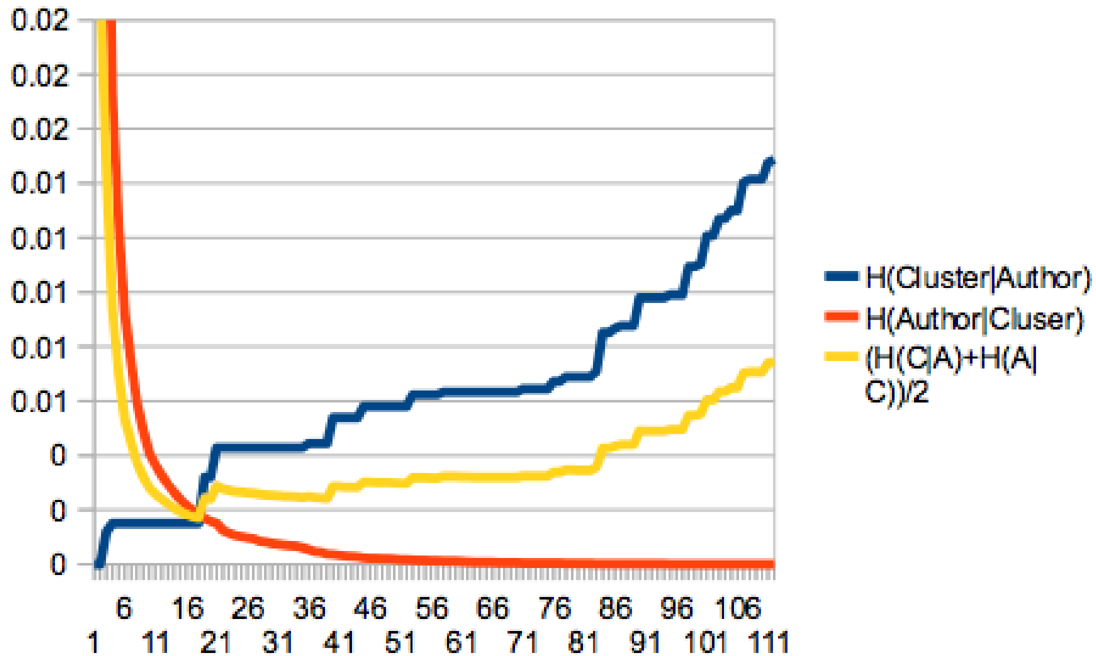
Clustering Properties
Model Selection - III

- Similar formulation using point-wise conditional entropy

$$H_{pointwise}(Y|X) \stackrel{\text{def}}{=} -\sum_{C} p(x,y)\log p(y|x)$$

$$C = \{c|c \in Clusters_{Co-Author}\}$$

- Example



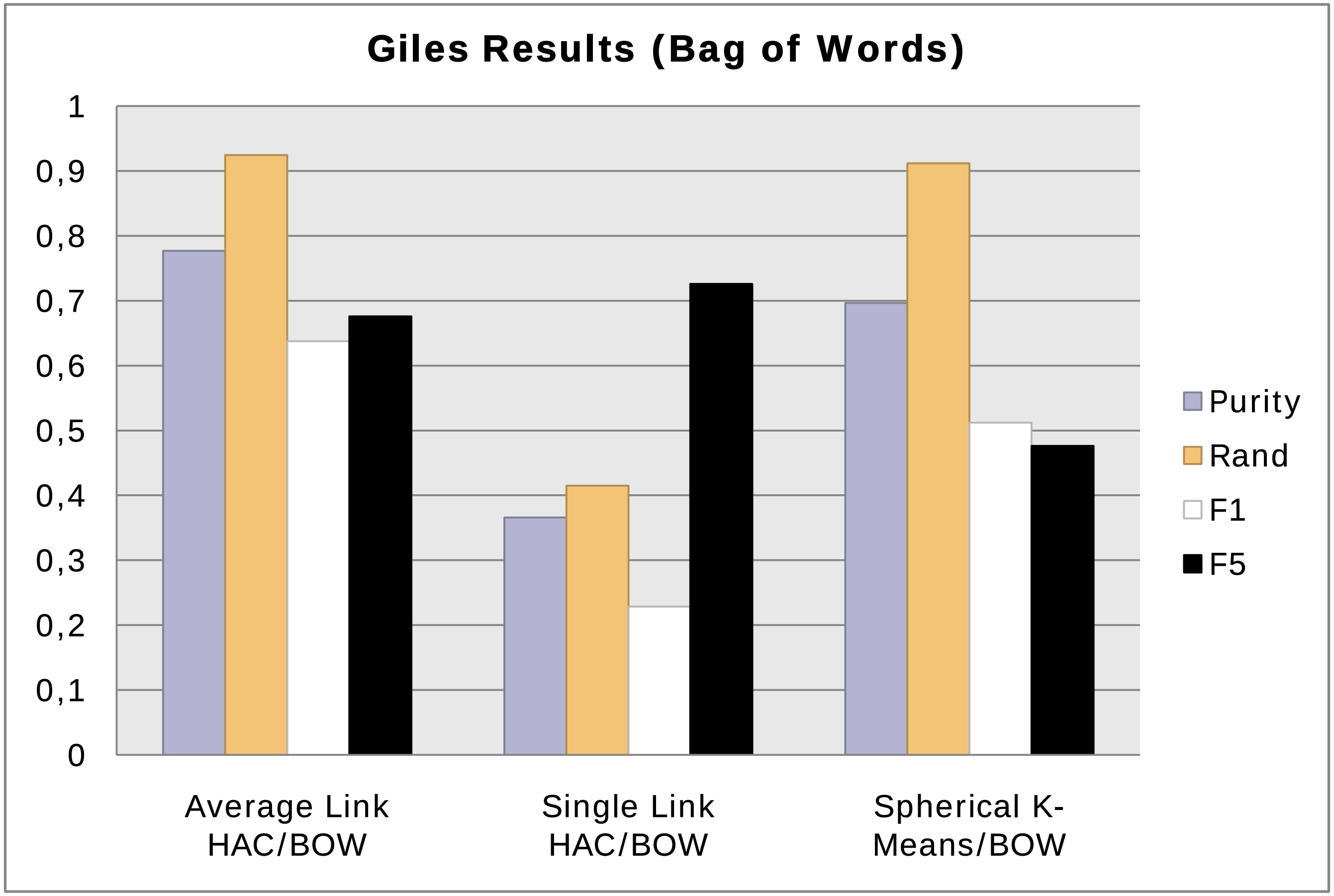
Experiments
Setup

- Two data set
  - **Giles** provides a nice dataset of citations with 12 ambiguous author names (http://bit.ly/aBV8qP)
  - **Mendeley** provided us with a much larger dataset retrieved from user profiles (<3 Mendeley)
  - For every publication/citation we also **gathered web search results** for additional data from Google, Bing, ACM (until we got blocked), e.g. plain text
- Workflow Setup
  - Identification (Step 1) was not necessary
  - Blocking – used Ground-Truth to create forename subsets: Lee, Martin, Gupta, Kumar, Chen, Johnson

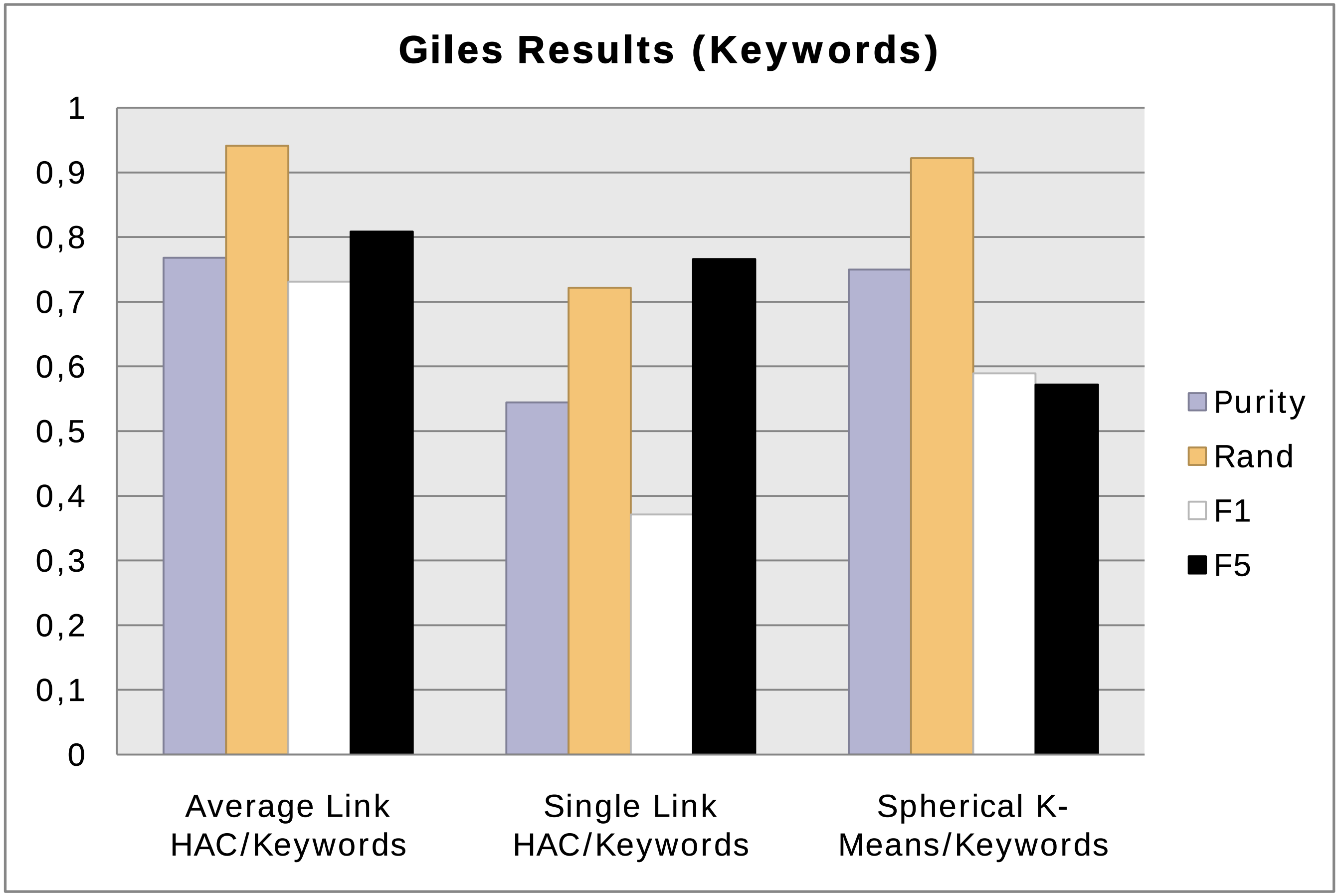
➔ **Error Analysis focuses solely on clustering properties**

Experiments
Results Clustering Algorithms

Fix number of clusters to the true number of unique authors



Experiments
Results Clustering Algorithms



HAC with average linking seems is best clustering approach

Experiments
Results Features

- Again, assume number of unique authors known

Author	Title	Keyword	Stem	Normalize	Purity	F1
✓	✓	✓		✓	0.92	0.9
✓	✓	✓	✓	✓	0.91	0.89
		✓		✓	0.87	0.84
		✓			0.87	0.84
✓		✓	✓	✓	0.88	0.83

Table I
BEST 5 RESULTS USING HAC CLUSTERING ON THE GILES-MARTIN SUBSET. SORTED BY F1. 16 DISTINCT AUTHORS, 112 PUBLICATIONS.

Author	Title	Keyword	Stem	Normalize	Purity	F1
✓	✓	✓	✓	✓	0.72	0.50
✓	✓	✓		✓	0.70	0.48
✓	✓	✓	✓		0.70	0.46
	✓	✓	✓	✓	0.61	0.42
✓	✓	✓			0.66	0.39

Table II
BEST 5 RESULTS ON HAC CLUSTERING ON THE GILES LEE SUBSET. SORTED BY F1. 100 DISTINCT AUTHORS, 1419 PUBLICATIONS.

- Results on the Martin subset are encouraging. Reason: full-text features contain less noise

Experiments

Results Model Selection

- What is the difference if we have to guess the author number?

Dataset	K_{real}	$F1_{best}$	K_{guess}	$F1_{guess}$	$F1_{real}$
Mendely-lee	49	61%	44	28%	27%
Giles-martin	16	90%	16	84%	84%
Giles-gupta	26	65%	14	43%	65%
Giles-kumar	14	70%	14	44%	44%
Giles-chen	61	46%	12	10%	37%
Giles-johnson	15	78%	11	60%	75%
Giles-lee	100	50%	21	5%	38%

Table IV
MODEL SELECTION RESULTS USING POINT-WISE CONDITIONAL ENTROPY ON KEYWORDS ONLY

- Performance varies, but gives good results when clustering comes close to the real groups (i.e. Martin Subset)
- Underestimate correct number of clusters

Conclusion

- HAC as empirical best algorithm for disambiguation
- New Model Selection Strategies work good given good clustering results
- Automatic Author Disambiguation still unsovled for practical scenarios
- Identification and Blocking as additional error sources
- Future Work
  - reduce the effect of blocking errors and model selection through outlier detection
  - Improved feature selection and cleaning

www.know-center.at

Thanks for your attention

and also thanks also to our supporters

http://mendeley.com

FP 7 TEAM Project: http://team-project.tugraz.at/

You can follow our research on http://team-project.tugraz.at/blog or
via Twitter @mgrani or
via Mendley‘s public group „Publications (TEAM IAPP)“