Erasmus Mundus European Master in
Language & Communication Technologies (LCT)

Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen - Bolzano

Nancy-Université
Université Nancy 2

Named Entity Disambiguation in Digital Libraries

Le Dieu Thu

Supervisors: Raffaella Bernardi
Massimo Poesio
Patrick Blackburn

Outline

1
- **Introduction**

2
- **Author Name Disambiguation: State of the art**

3
- **Our Disambiguation Framework**
  - Blocking module
  - Metadata Enriching module
  - Feature representation & cosine similarity
  - Sparsity problem & Dimensionality Reduction
  - Clustering module

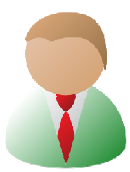
4
- **Experiments & Discussion**

5
- **Conclusions**

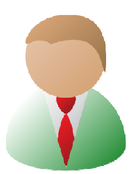
Introduction

- **Ambiguous author name:** Different authors having the same name
- **Aim:** Disambiguate those authors

R Smith → Author of → Books



- High finance in the Euro zone : competing in the new European capital market
- Governing the modern corporation : capital markets, corporate control and economic performance
- The money wars : the rise and fall of the great buyout boom of the 1980s



- Ecology and field biology
- Elements of ecology

Bolzano Library Catalogue Searching

- **Classification Numbers:** encodings (QK 620, WI 2000), to organize books on shelves
- **Subject Headings:** keywords

High finance in the Euro zone : competing in the new European capital market

QK 620

Finance, Economics

Ecology and field biology

WI 2000

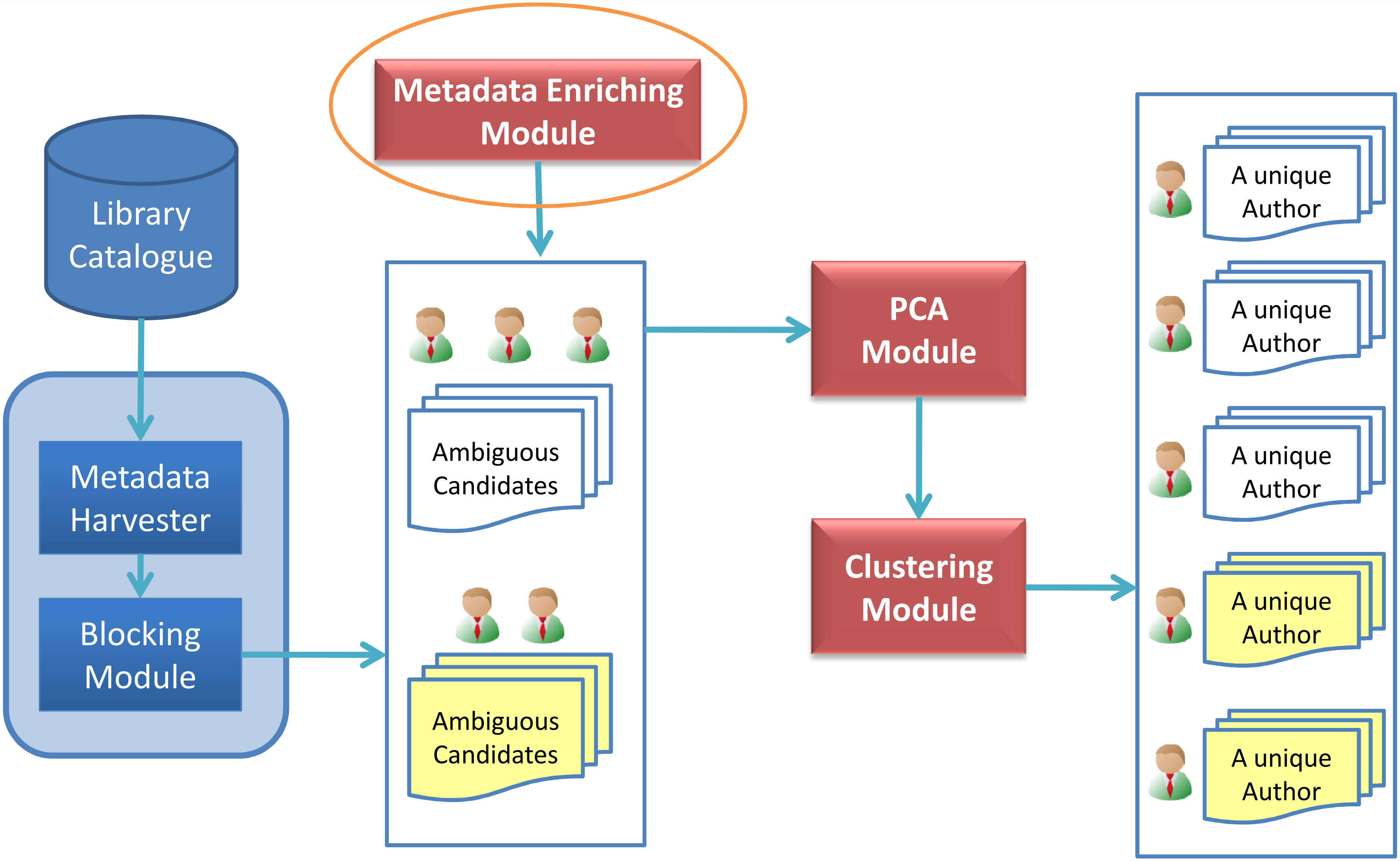
Ecology, Biology

Searching by *CN* & *SH*:

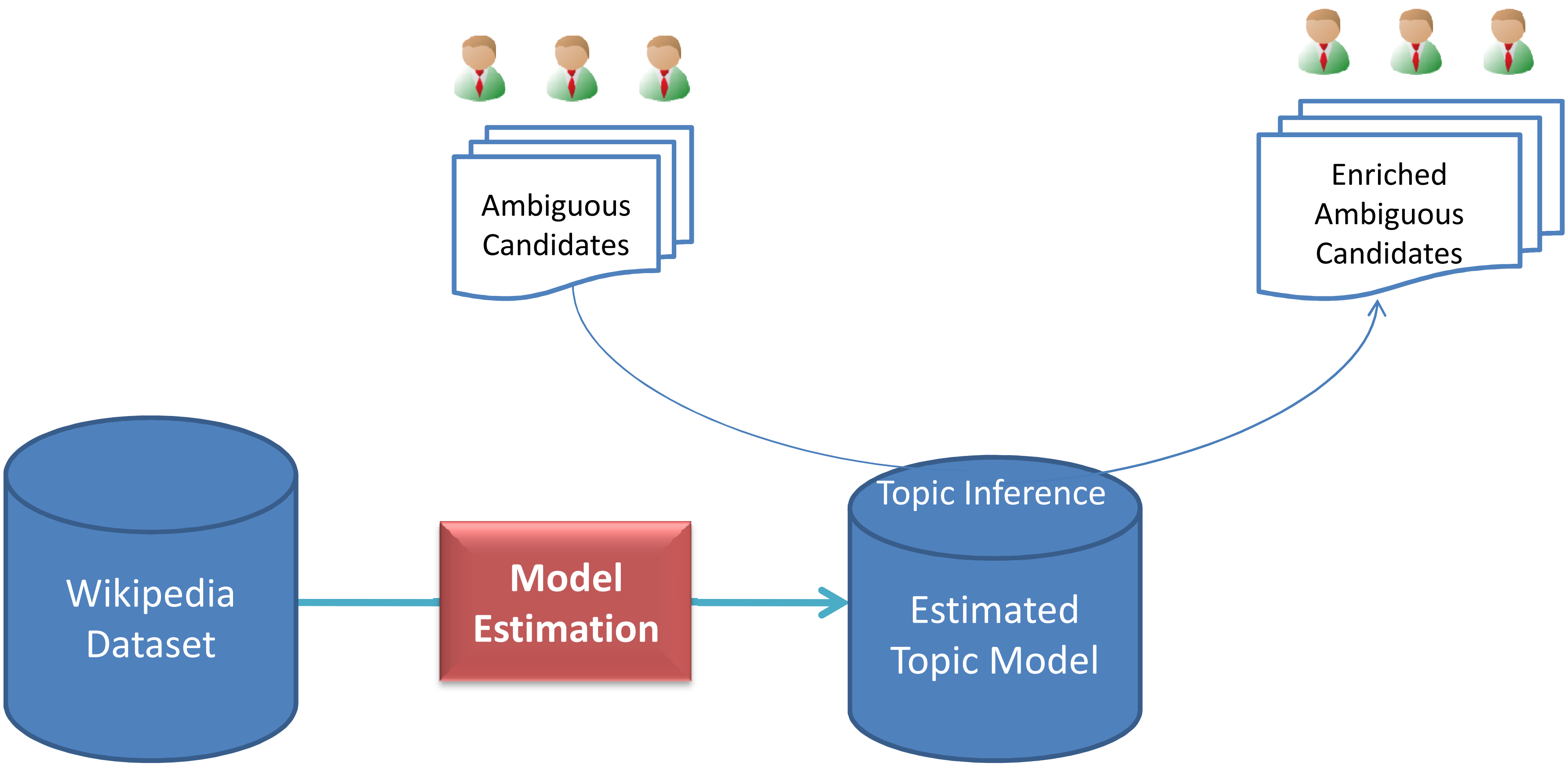
Not of support in Digital Libraries (CiteSeer, DBLP)

Goal: Develop a disambiguation method that can be applied in any catalogue (even without manual annotations, e.g. *CN* & *SH*)

Proposed Disambiguation Framework



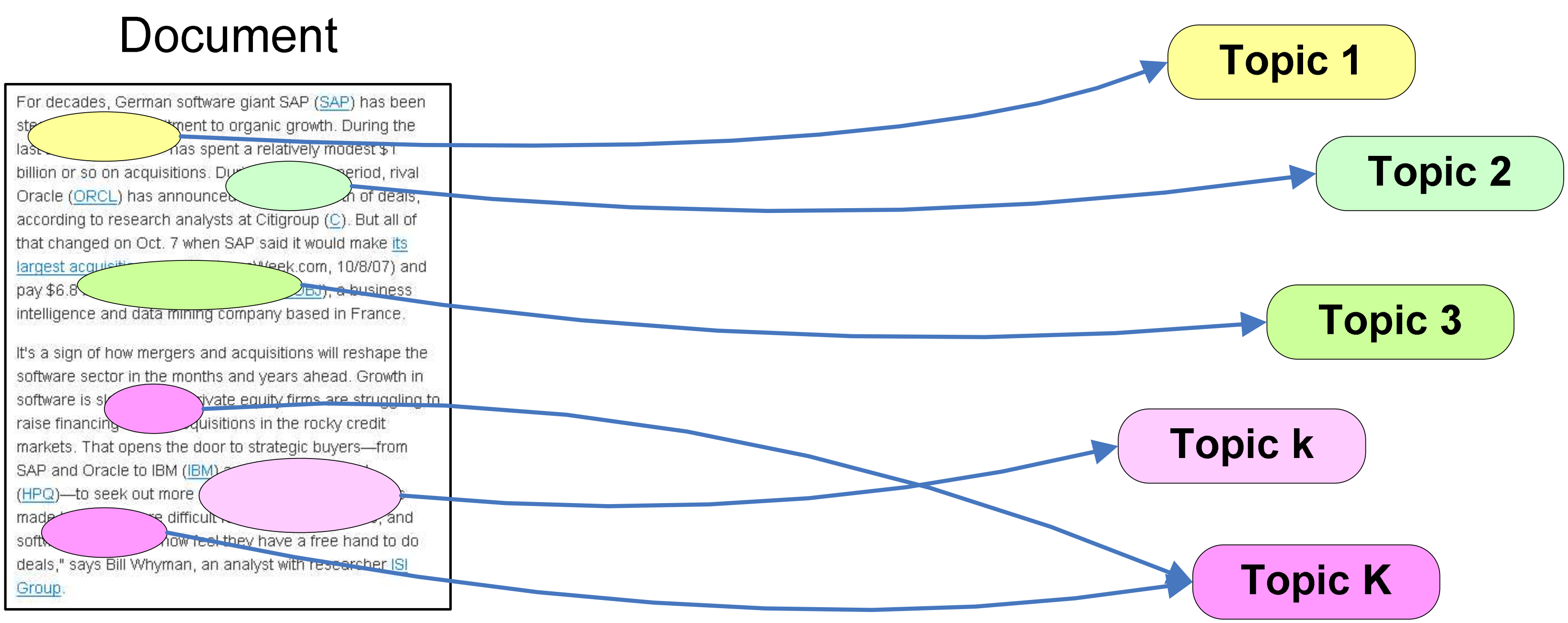
Metadata Enriching Module



Topic Models:

Hidden Topic Discovery from Documents

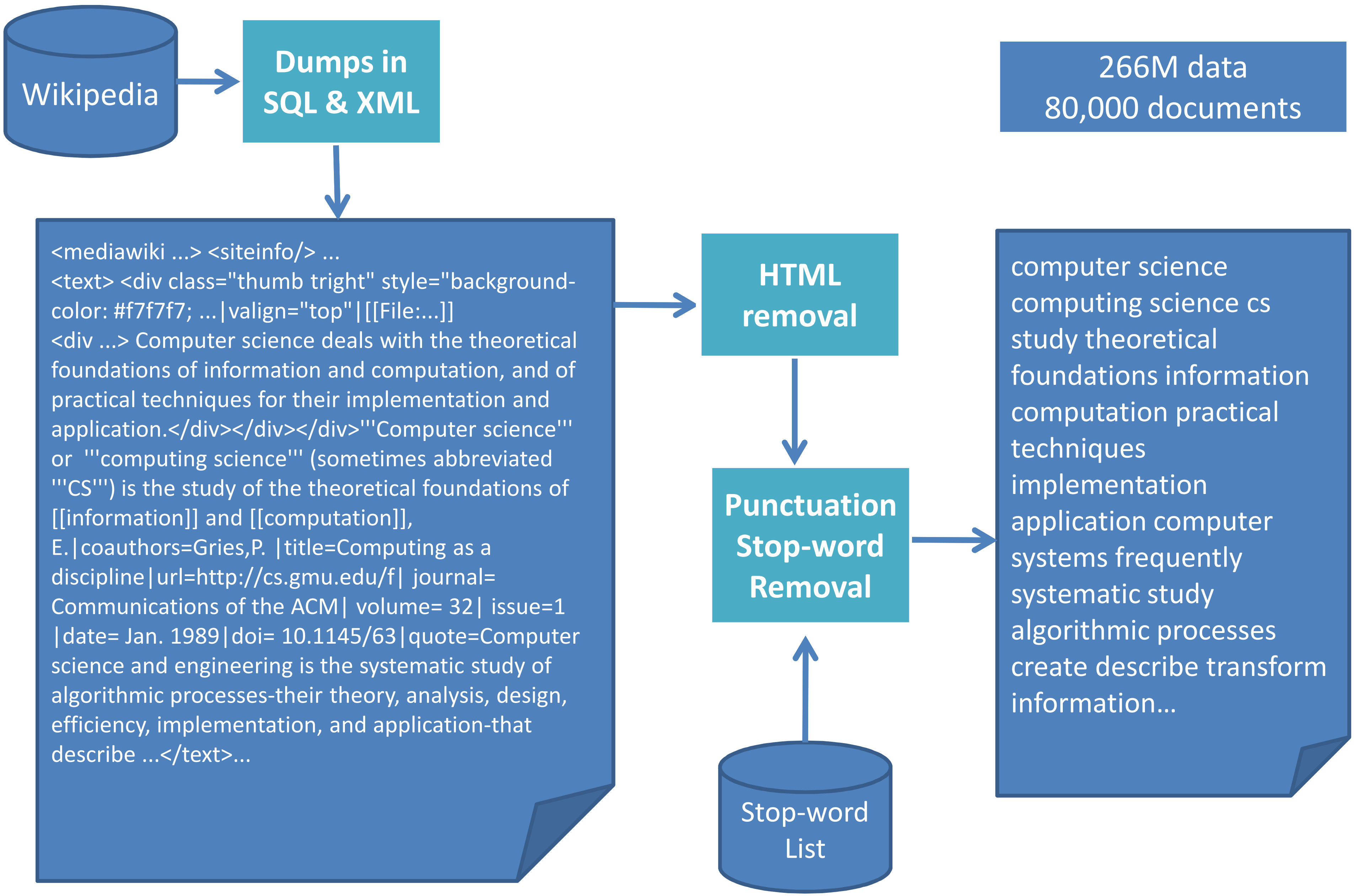
Latent Dirichlet Allocation (**LDA**) [Blei et al. 2003]



Hidden Topic Analysis/Discovery

Topics {1, 2, …, K} are unknown (i.e., hidden and need to be discovered)

Wikipedia data preprocessing



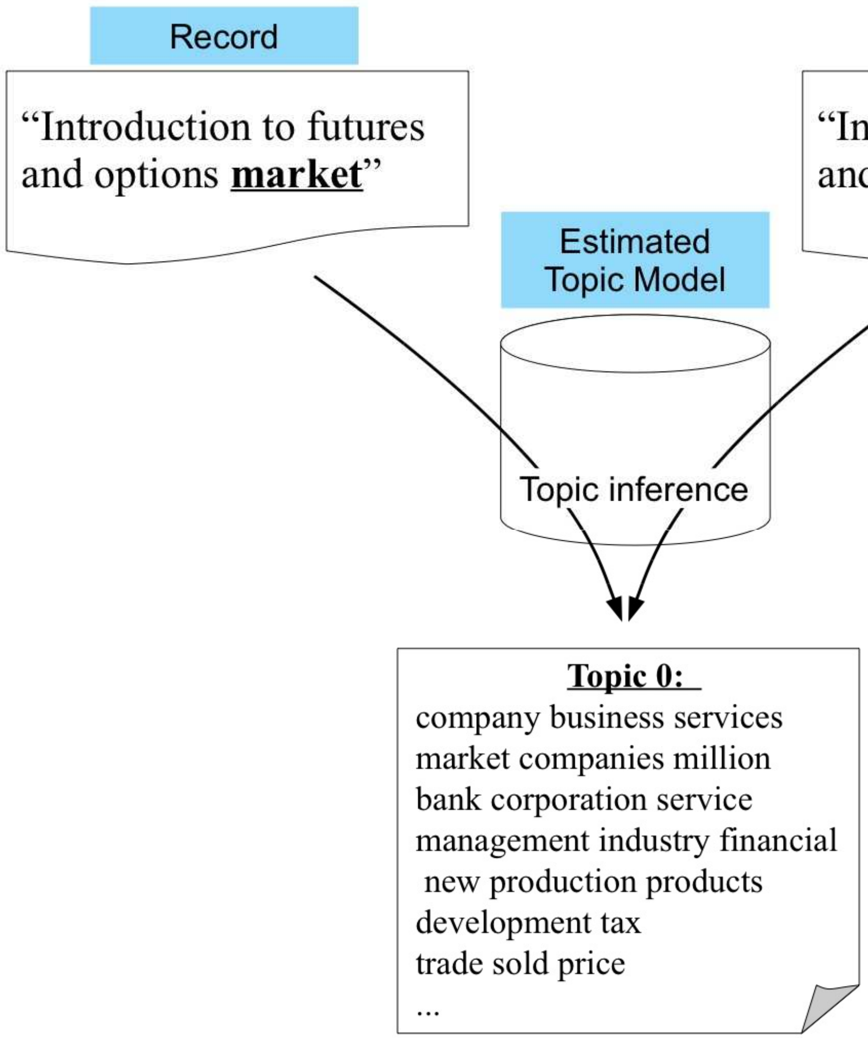
Model Estimation using LDA & Gibbs Sampling

Sample topics extracted from the estimated model

Topic 0	Topic 8	Topic 23	Topic 39	Topic 68	Topic 86	Topic 96
company	album	cells	law	storm	war	school
business	music	disease	court	tropical	army	university
services	band	medical	police	damage	force	college
market	song	patients	legal	winds	battle	high
companies	released	treatment	rights	typhoon	military	students
million	singer	cell	public	cyclone	air	schools
bank	rock	blood	justice	storms	navy	education
service	guitar	health	laws	caused	ship	institute
industry	live	medicine	judge	landfall	command	year
financial	records	brain	criminal	season	attack	program
tax	vocals	protein	supreme	pacific	fire	campus

Toolkit: GibbsLDA++; 1000 iterations; 2.8GHz computer; Heap size: 3G; took 14 hours

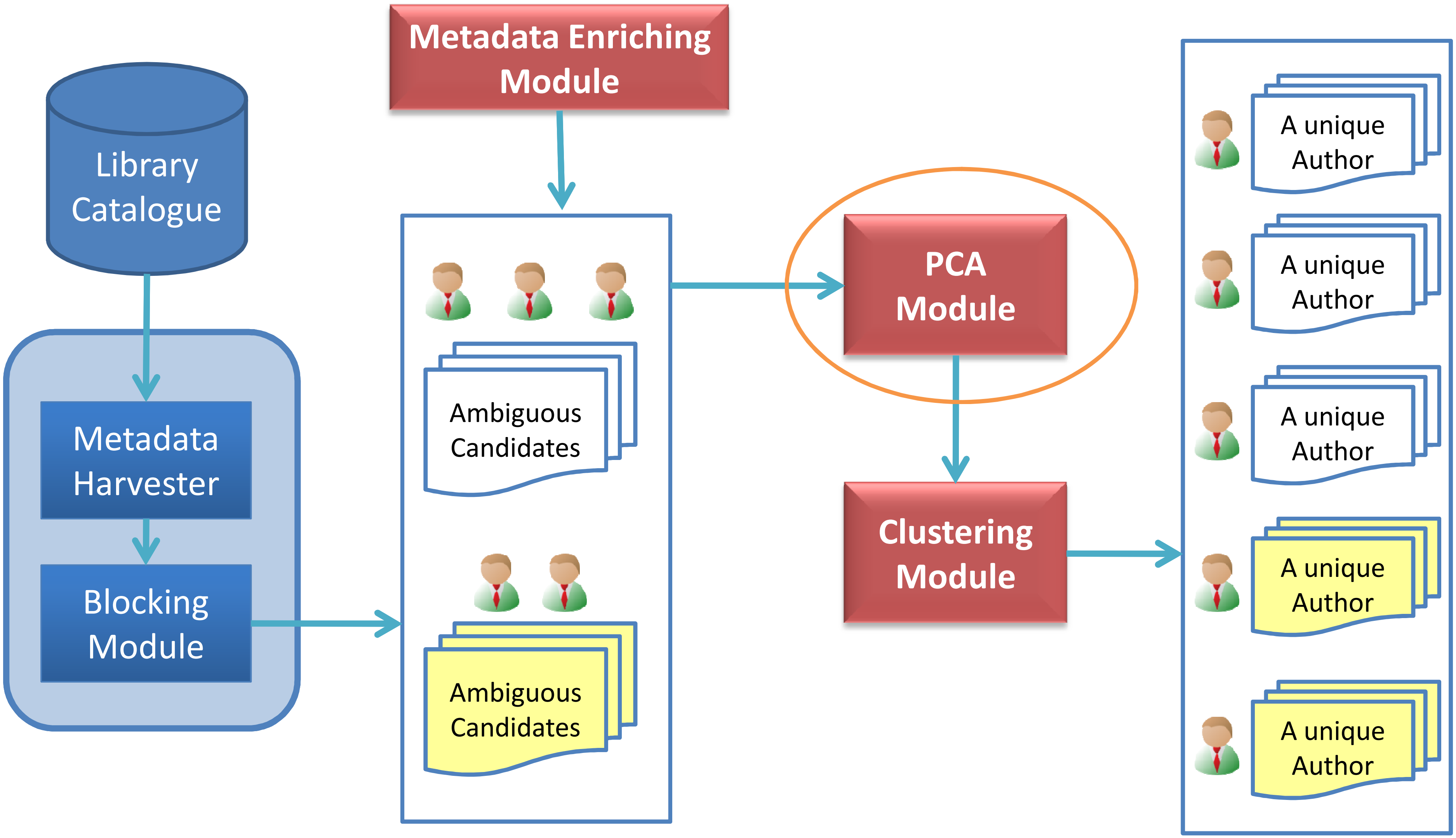
Hidden Topic Inference for Metadata



probability of topic *k* in record *m*

$$Frequency_{m,k} = \begin{cases} round\left(scale \times \vartheta_{m,k}\right), \text{if } \vartheta_{m,k} \geq cut\text{-}off \\ 0, \text{if } \vartheta_{m,k} < cut\text{-}off \end{cases}$$

Proposed Disambiguation Framework

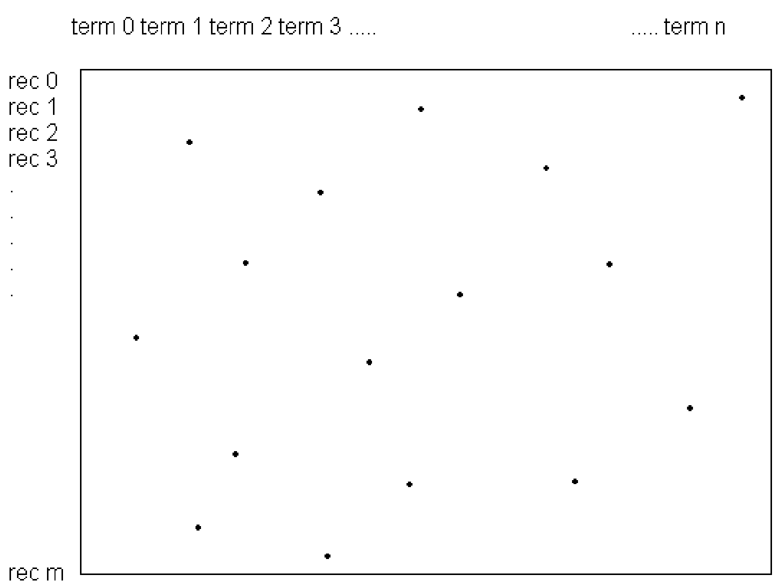


Feature Representation

- Features: (co-author names, title, publishers)
- Feature Representation: Vector Space Model
- Record similarity: Cosine similarity

weight of t_i

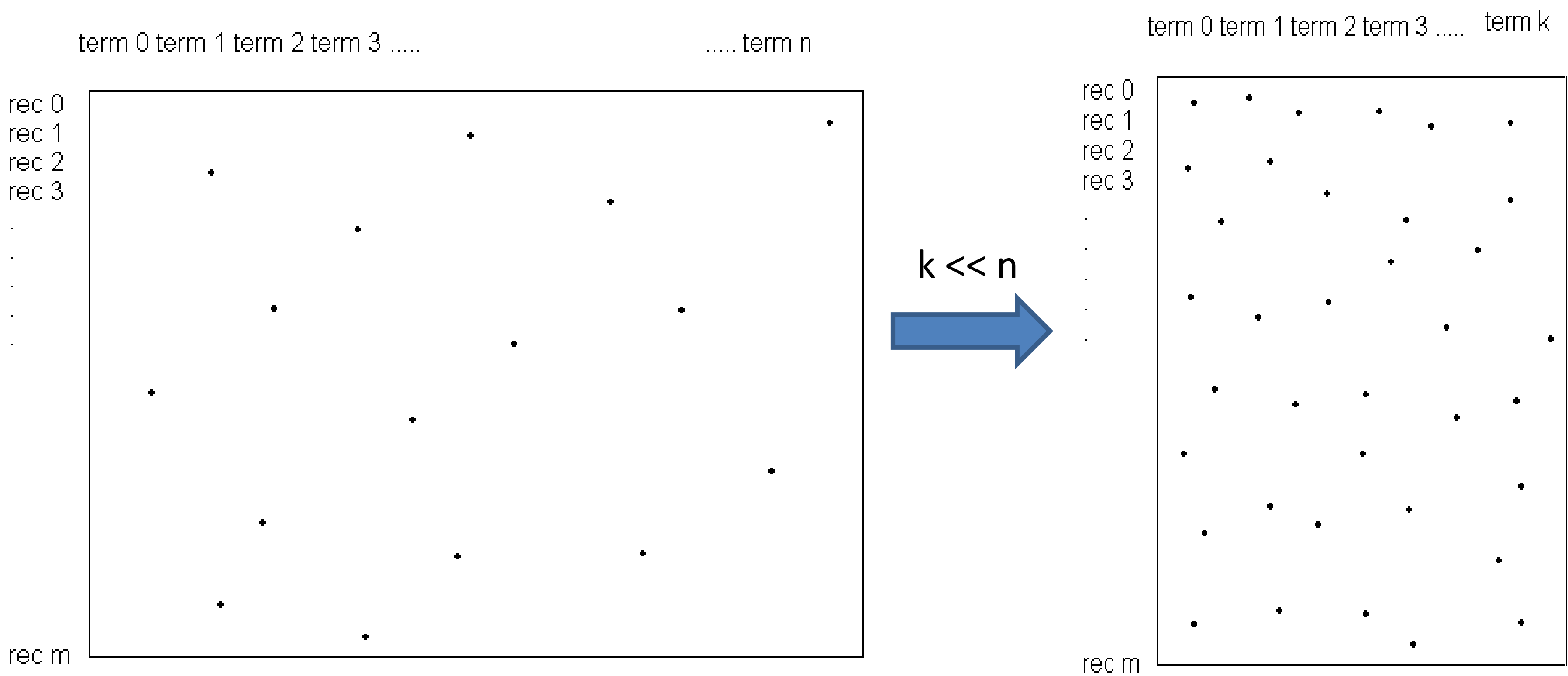
$$cosin\_sim(\mathbf{r}_i, \mathbf{r}_j) = \frac{\vec{\mathbf{r}_i}.\vec{\mathbf{r}_j}}{|\vec{\mathbf{r}_i}|.|\vec{\mathbf{r}_j}|} = \frac{\sum_{t \in V} w_{ti}.w_{tj}}{\sqrt{\sum_{t \in V} w_{ti}^2}.\sqrt{\sum_{t \in V} w_{tj}^2}}$$



High Dimensional Space

Sparseness

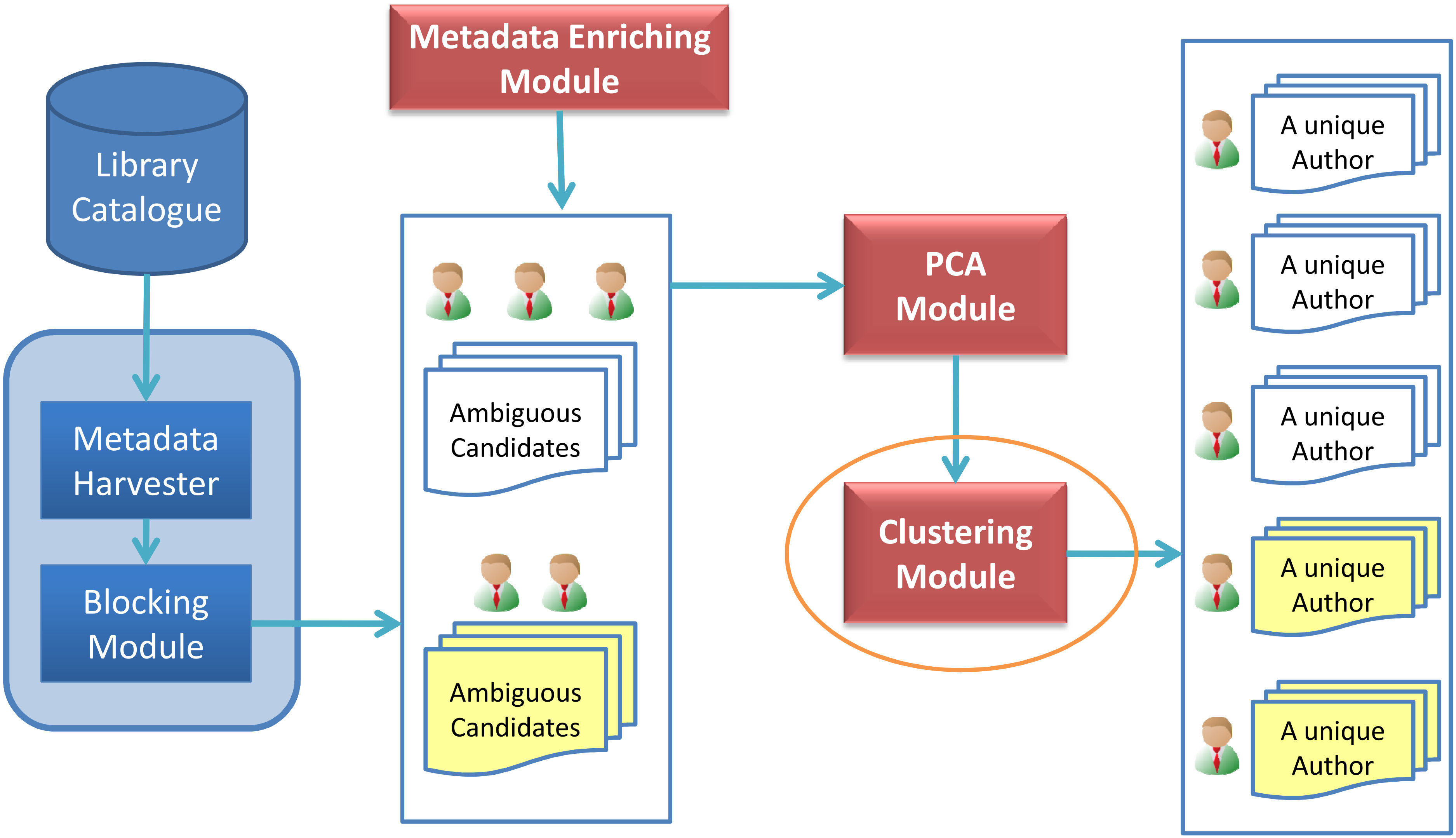
Dimensionality Reduction with PCA



Principle Component Analysis (PCA): reduce each vector to few dimensions while keeping as much of the variance as possible.

- Less sparse
- More understandable model (visualization for better quantity analyses)
- Reduce speed & complexity of the Clustering process

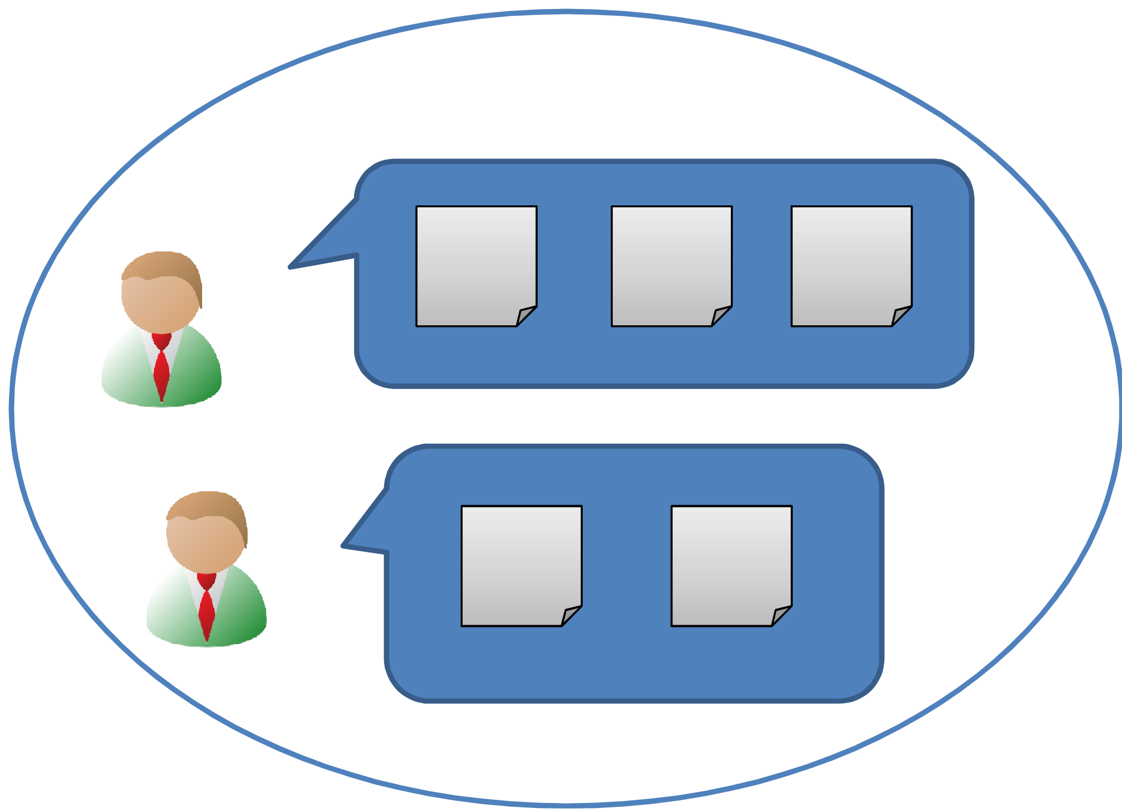
Proposed Disambiguation Framework



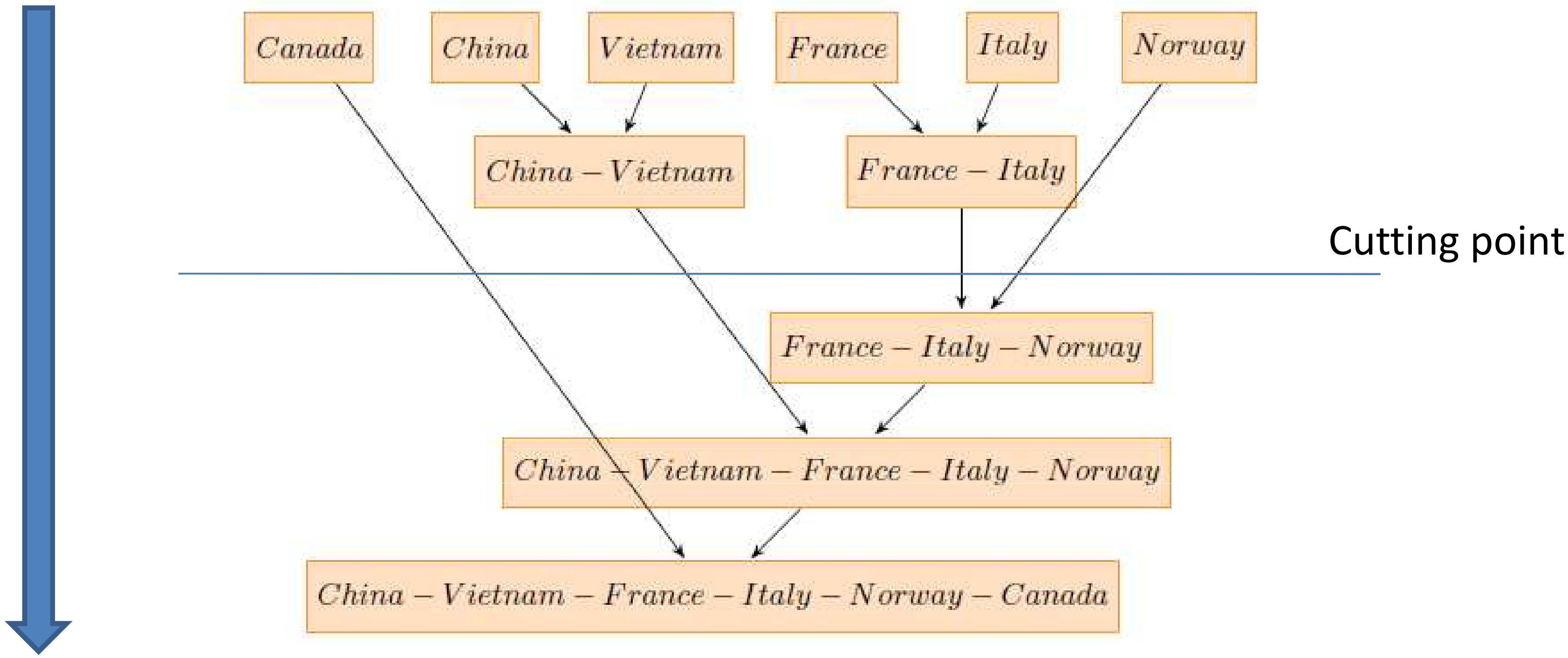
Clustering Module

Author Name Disambiguation ⟷ Clustering Problem

- **Cluster books:** each cluster corresponds to a unique author
- Which clustering algorithm?
  - Number of clusters?
  - Distance metric?



Clustering Module: HAC Algorithm



- Clustering algorithm: Hierarchical Agglomerative Clustering (HAC)
- Distance between Clusters (A & B): Average Linkage

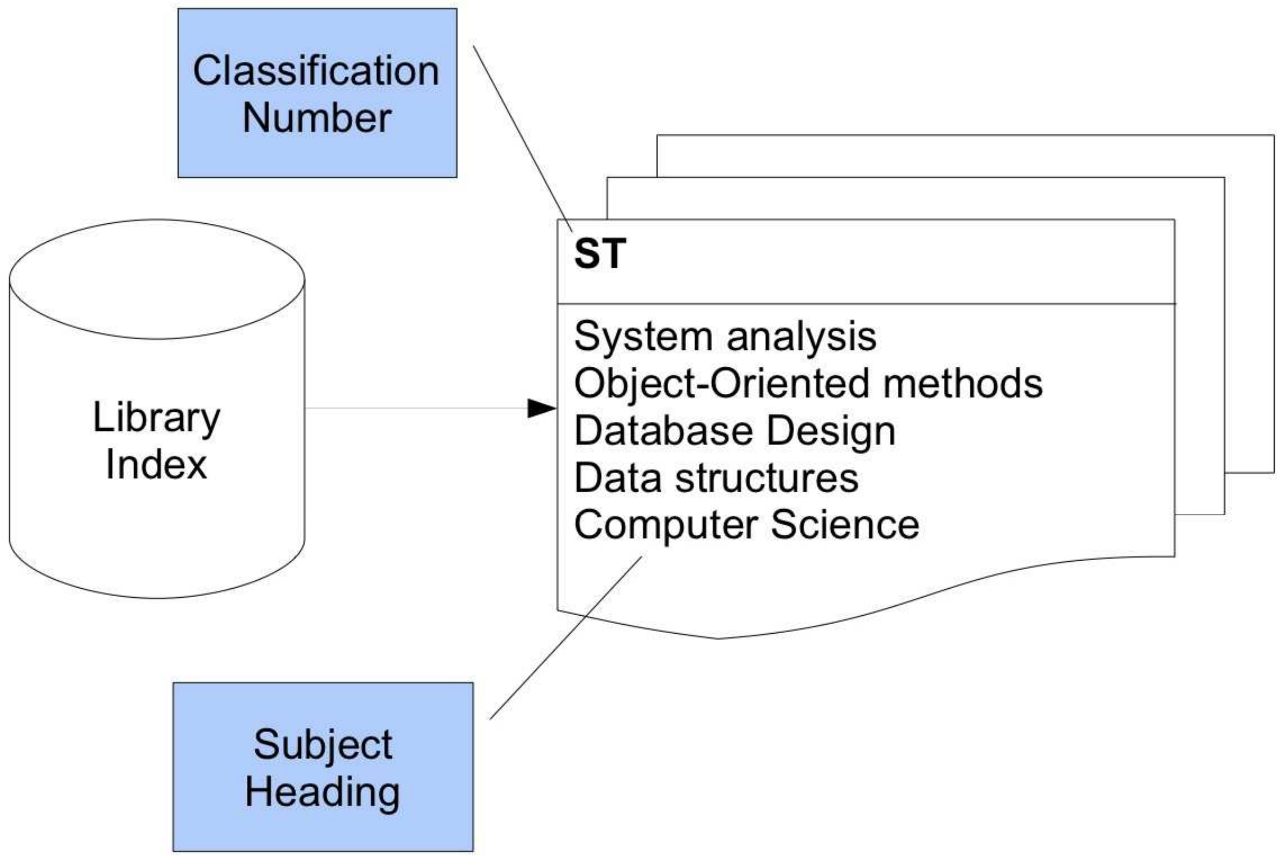
$$d(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{A}|.|\mathcal{B}|}.\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} d(x,y)$$

Experimental Settings

Settings	Records
Baseline	$c \cup t \cup p$
CNSH	$(c \cup t \cup p) \oplus (CN \cup SH)$
CNSH-Enriched	$(c \cup t \cup p) \oplus (CN \cup SH) \oplus$ *SH-enriched*
CNSH-PCA	$[(c \cup t \cup p) \oplus (CN \cup SH)]_{PCA}$
HT	$(c \cup t \cup p) \oplus HT$

- $c =$ co-author names
- $t =$ book's title
- $p =$ book's publisher
- $CN =$ book's Classification Numbers
- $SH =$ book's Subject Headings
- *SH-Enriched*: Set of Enriched Subject Headings
- PCA: applying PCA to reduce dimensions
- HT: Set of most likely hidden topics inferred from the estimated topic model

SH-Enriched



- **Goal:** exploit as much as possible the manual annotation information (i.e., *CN, SH*)
- Extract all English books (29,000 books): group *SH* by *CN*

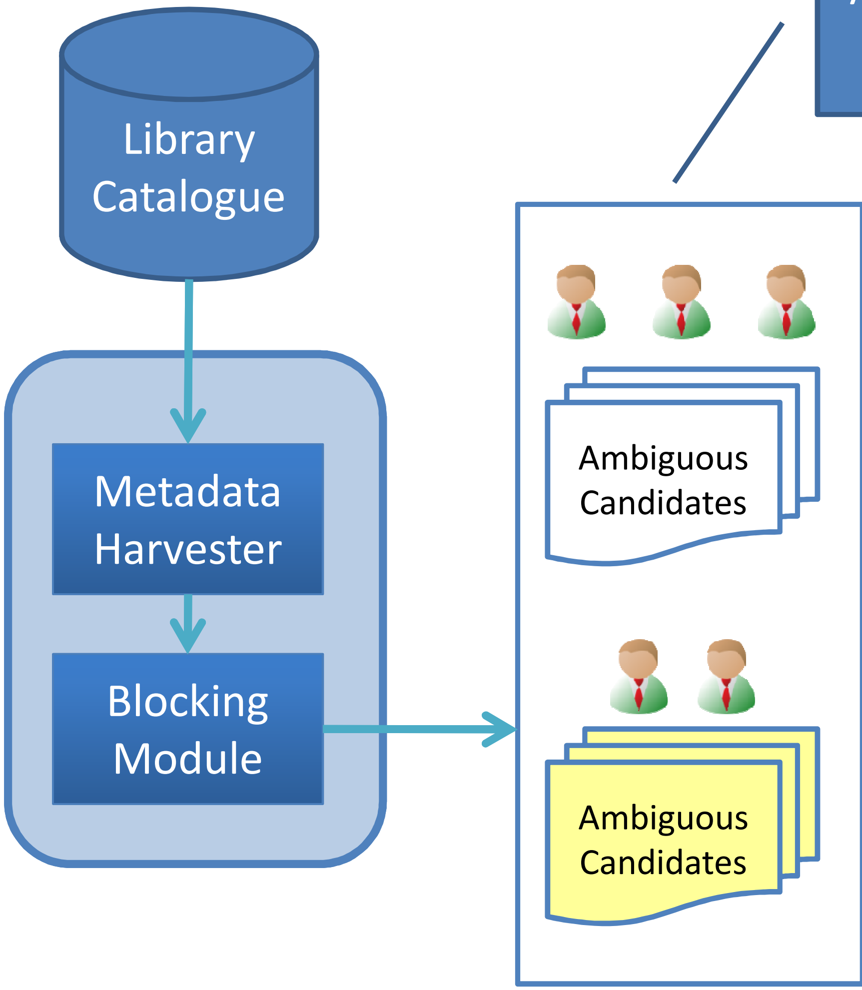
SH-Enriched

Classification Number	CC	DK	ET	ST	WF
Subject Headings	religion	observation	phonetics	metadata	microbiology
	addresses	research	grammar	PHP	organisms
	essays	project	lexicography	XSLT	soil
	lectures	education	philosophy	computer	food industry
	civilization	school	semantics	program	crops
	philosophy	program	cognition	language	nitrogen
	ethics	learning	phonology	software	microbial
	cognition	daycare	typology	Microsoft	innovations
	evolution	child development	linguistics	design	government policy

- Each record is enriched with the **first 20** *SHs* in the corresponding *CN*

Experimental Data

Authors sharing the first initial name & last name (e.g., R Smith) **[Torvik 05, 09; Huang 06]**



- **Data for testing from Bolzano Library catalogue**
  - 28 groups of authors (each group contains 2 – 30 distinguished authors)
  - books in English, having full information

  → **Preprocessing:** normalization, segmentation, stop-word removal

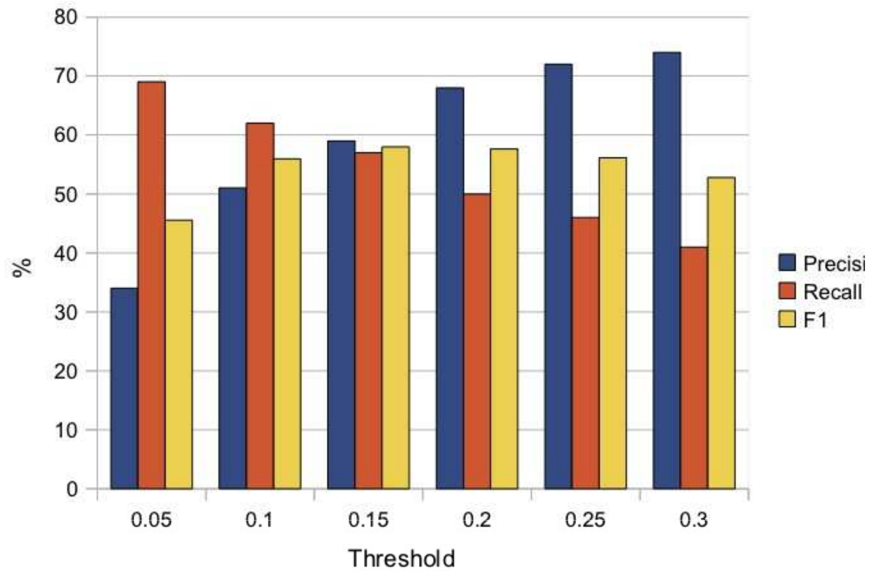
Evaluation Metrics

$$pPre = \frac{number\ of\ correct\ pairs\ in\ the\ output\ clusters}{number\ of\ total\ pairs\ in\ the\ output\ clusters}$$

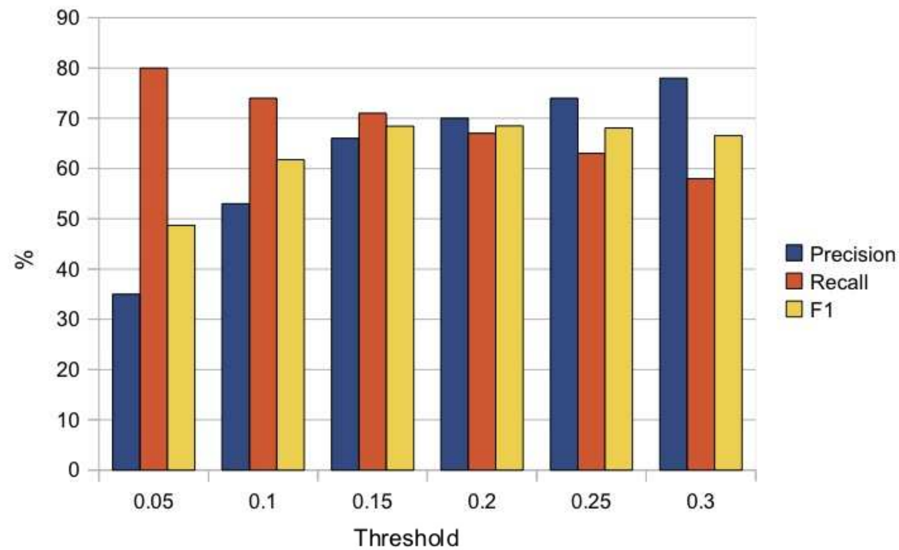
$$pRe = \frac{number\ of\ correct\ pairs\ in\ the\ output\ clusters}{number\ of\ total\ pairs\ in\ the\ truth\ clusters}$$

$$pF1 = 2.\frac{pPre.pRe}{pPre + pRe}$$

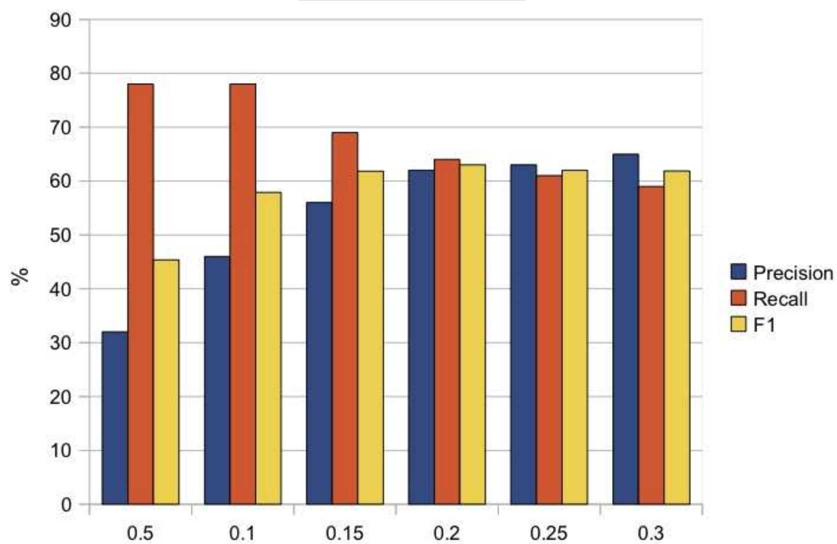
Results



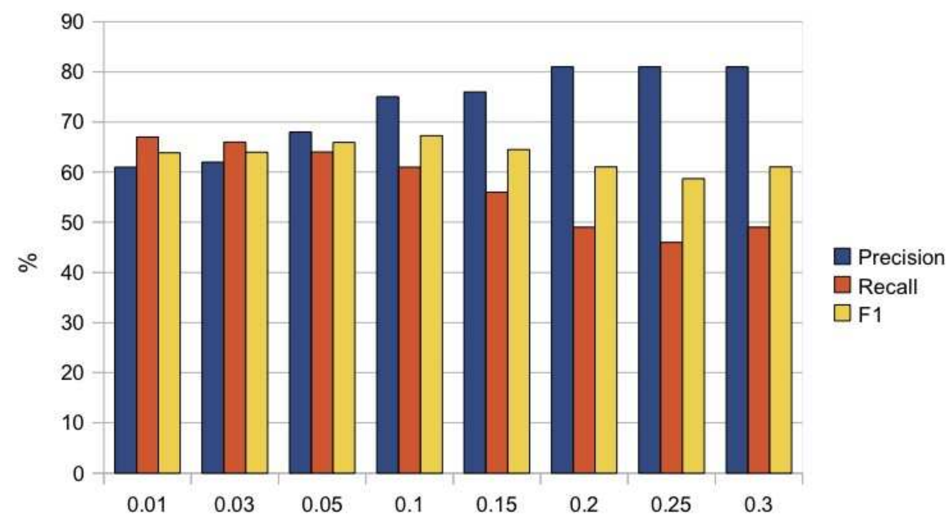
Baseline



CNSH

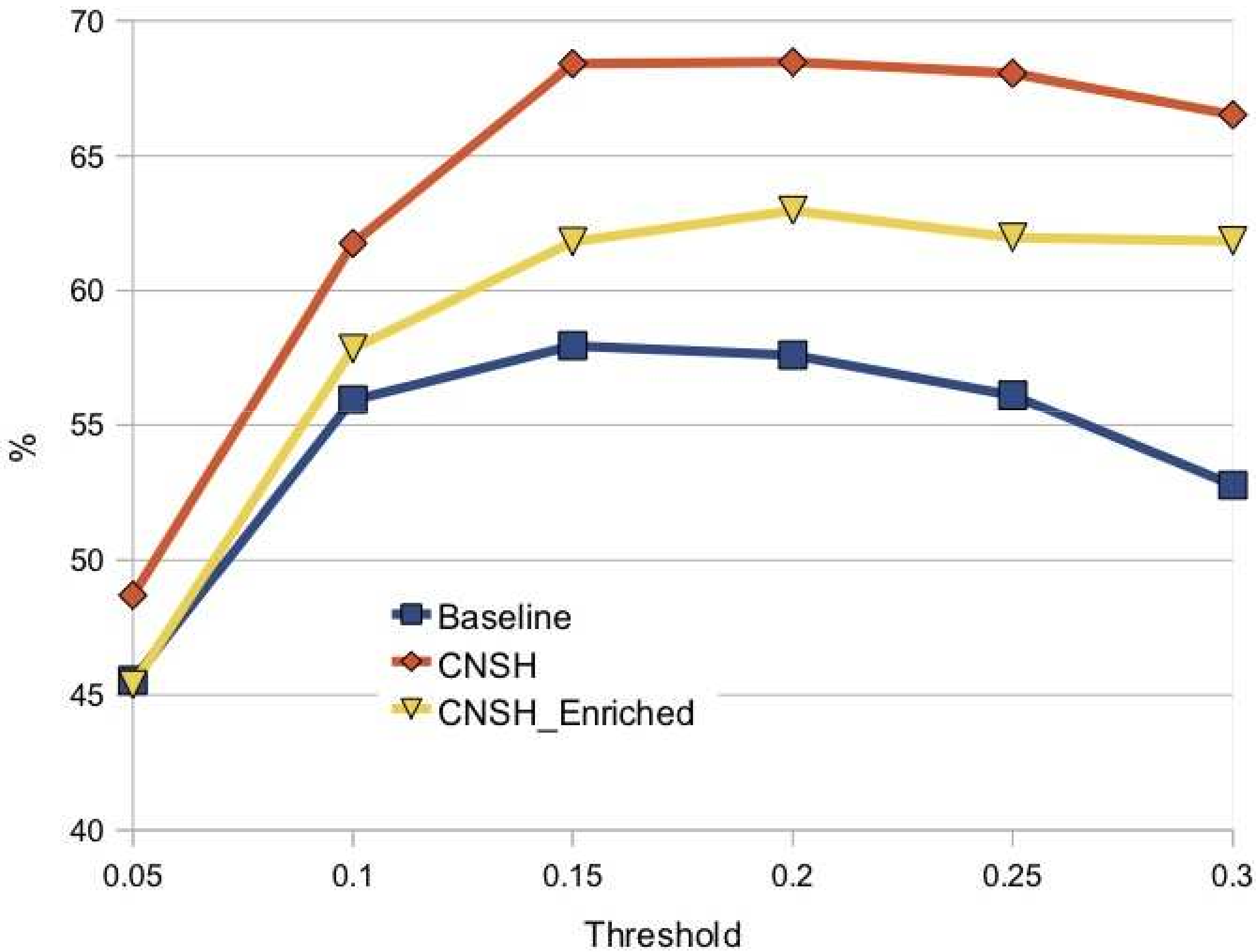


CNSH-Enriched

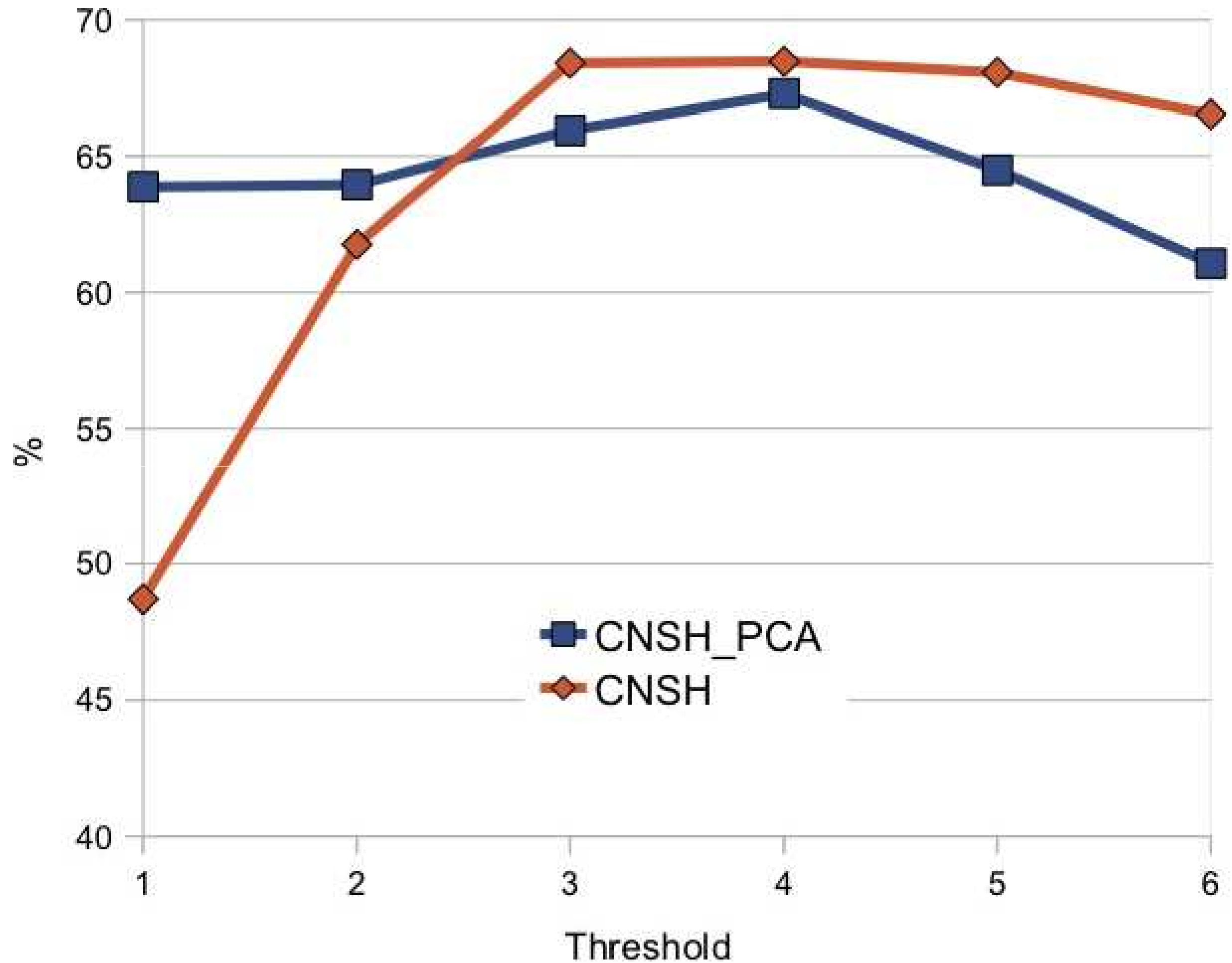


CNSH-PCA

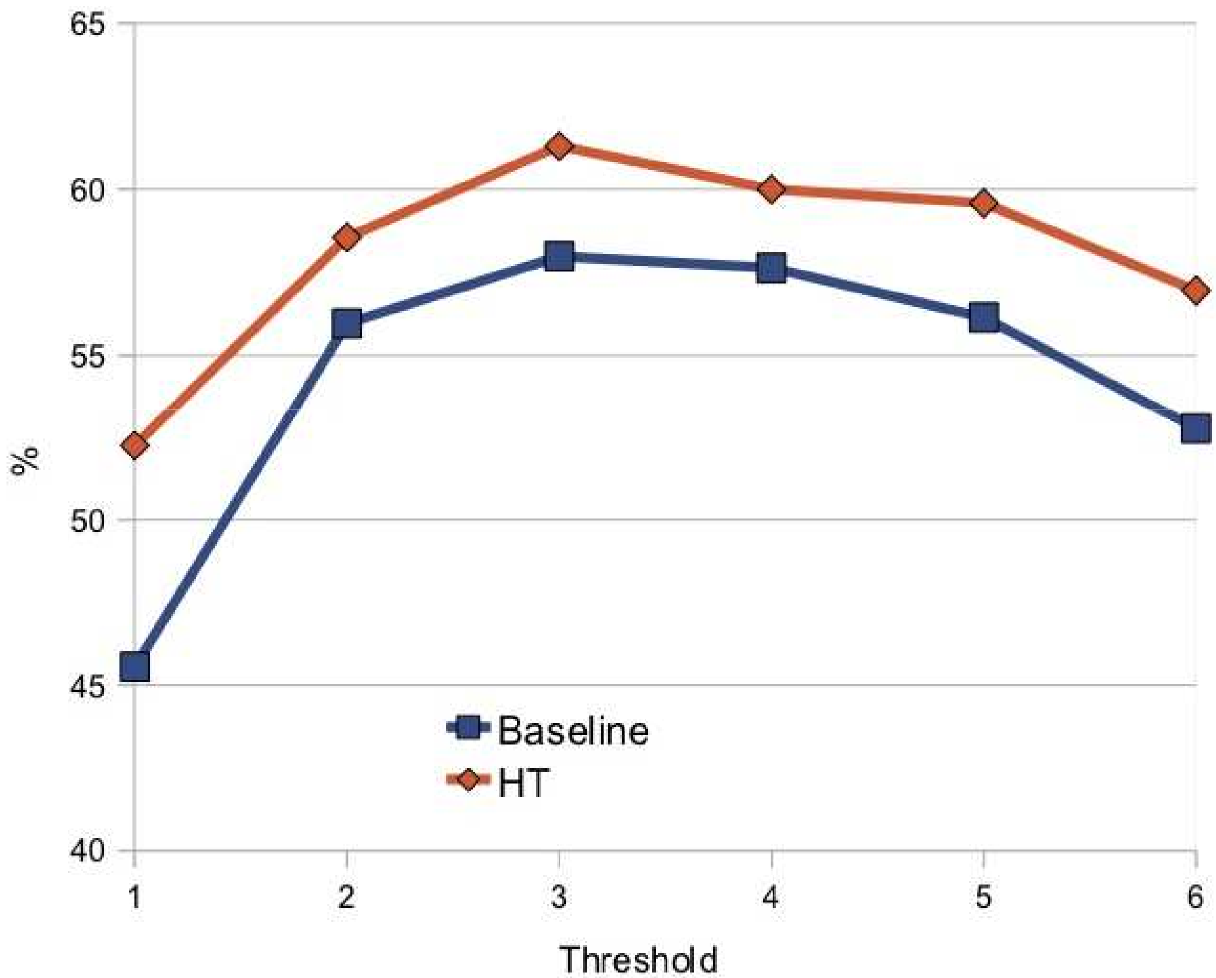
CNSH vs. CNSH-Enriched



CNSH vs. CNSH-PCA



Hidden Topic Enriching performance



Conclusions

- Proposed a framework for author name disambiguation:
  - exploit manual annotations (*CN, SH*)
  - use Hidden Topics estimated from Wikipedia to automatically enrich record's information
    - can be applied to federated libraries, digital libraries like CiteSeer, DBLP, PubMed
    - take advantage of available large-scale knowledge-base dataset, Wikipedia
    - can be used for different languages
  - use PCA to represent data in a more compact way
    - reduce number of dimensions → reduce speed & complexity, reduce noises
    - achieve satisfactory results
    - can be used for visualization for better quantity clustering analyses in the future

Future works

- Contribution of different features
- Experiment in a multilingual environment
- Optimize cutting points for HAC

Acknowledgement

Raffaella Bernardi

Massimo Poesio

Patrick Blackburn

Luigi Siciliano

CACAO project

Marco Baroni

Cristiano Cumer

Manuel Kirschner

European Commision & LCT program

Main References