CREST

Big Data Application
(2015.10-2021.3)

July 11, 2018
DATAIA-JST Symposium

# Scientific Paper Analysis:

## Knowledge Discovery through Structural Document Understanding

Yuji Matsumoto
Nara Institute of Science and Technology
Nara, Japan

NAIST

# Objective of the Project

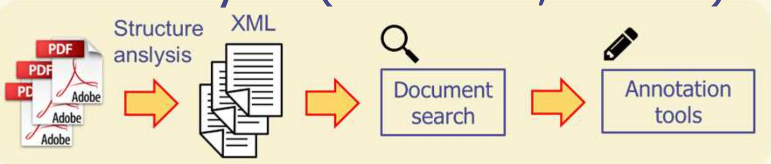
- Processing scholarly documents (scientific papers) to help researchers
  - To find similar/dissimilar research papers
  - To grasp contents of papers
  - To extract domain knowledge
  - To visualize extracted information
  - To support decision/idea making
- Development of tools and environment for scientific paper analysis, visualization, and acquisition

# Research Groups

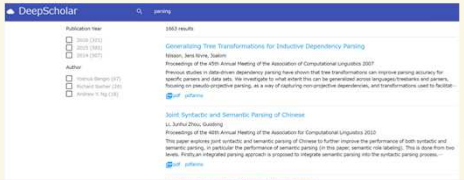
- G0: Matsumoto, Shindo, Shimbo, … (NAIST)
  - Semantic and structure analysis of scholarly documents
  - Knowledge extraction from scholarly documents
- G1: Satoh (NII), Nguyen (JAIST)
  - Legal text processing and information extraction
- G2: Inui, Inoue (Tohoku U)
  - Evidence mining in scientific documents
- G3: Aizawa, Miyao, Abekawa(NII),Nanba(Hiroshima City U)
  - Document analysis / Resource construction (ACL Anthology corpus)
- G4: Tsuruoka (U Tokyo)
  - Text summarization / Question answering in scientific fields
- G5: Mori (U Tokyo)
  - Citation Analysis: Detecting Research Trend of Academic Fields
- G6: Kano (Shizuoka U)
  - Brain map construction / visualization from table data extraction

# Tasks and Groups

## PDF Analysis (G0:NAIST, G3:NII)



## Document Search interface (G0, G1)



Publication Year

□ 2016 (321)
□ 2015 (583)
□ 2014 (507)

Author

□ Yoshua Bengio (37)
□ Richard Socher (28)
□ Andrew Y. Ng (18)

1663 results

Generalizing Tree Transformations for Inductive Dependency Parsing

Nilsson, Jens Nivre, Joakim

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics 2007

Previous studies in data-driven dependency parsing have shown that tree transformations can improve parsing accuracy for specific parsers and data sets. We investigate to what extent this can be generalized across languages/treebanks and parsers, focusing on pseudo-projective parsing, as a way of capturing non-projective dependencies, and transformations used to facilitat…

pdf pdfanno

Joint Syntactic and Semantic Parsing of Chinese

Li, Junhui Zhou, Guodong

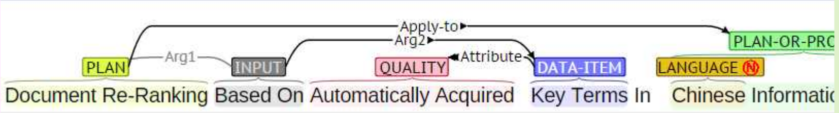
Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics 2010

This paper explores joint syntactic and semantic parsing of Chinese to further improve the performance of both syntactic and semantic parsing, in particular the performance of semantic parsing (in this paper, semantic role labeling). This is done from two levels. Firstly,an integrated parsing approach is proposed to integrate semantic parsing into the syntactic parsing process.…

pdf pdfanno

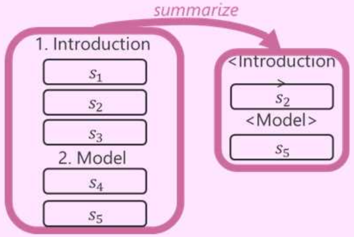
## Text Analysis

Named Entity / Relation analysis (G


Apply-to
Arg2
Arg1
Attribute
PLAN
INPUT
QUALITY
DATA-ITEM
LANGUAGE
PLAN-OR-PRO
Document Re-Ranking Based On Automatically Acquired Key Terms In Chinese Informatio

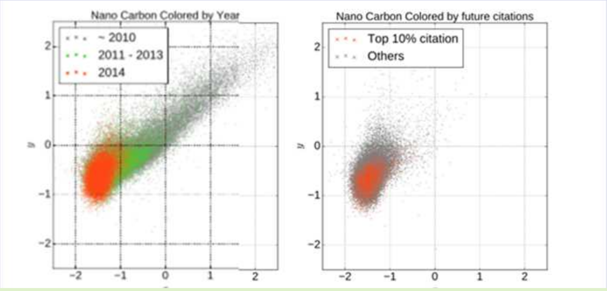
Logical structure analysis (G2:Toho

Document Summarization (G4:Tokyo)



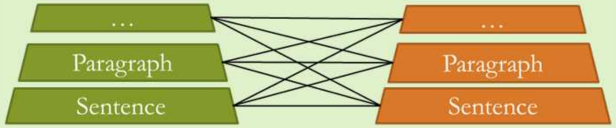
## Citation Graph Analysis

Trend prediction(G5:Tokyo)

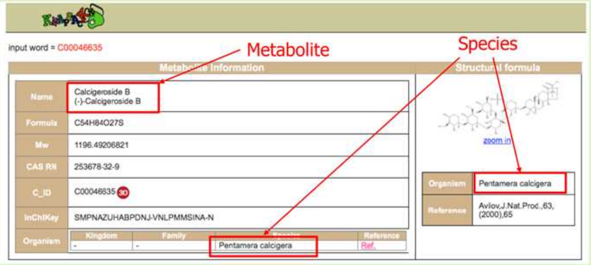


## Collaboration with domain experts

Legal text processing/search (G1:NII)



Domain KB acquistion (G0)



input word = C00046635

| Metabolite Information | |
|---|---|
| Name | Calcigeroside B<br>(-)-Calcigeroside B |
| Formula | C54H84O27S |
| Mw | 1196.49206821 |
| CAS RN | 253678-32-9 |
| C_ID | C00046635 3D |
| InChIKey | SMPNAZUHABPDNJ-VNLPMMSINA-N |

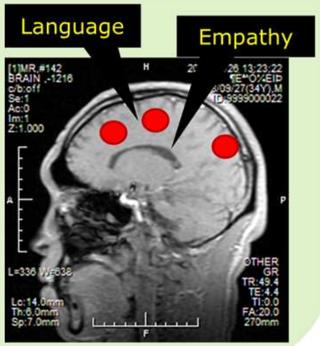
| Organism | Kingdom | Family | Species | Reference |
|---|---|---|---|---|
| | - | - | Pentamera calcigera | Ref. |

Structural formula

| | |
|---|---|
| Organism | Pentamera calcigera |
| Reference | Avilov,J.Nat.Prod.,63,(2000),65 |

KEGG pathway DB (G0)

Brain map (G6)



# Research Items for Scientific Document Analysis

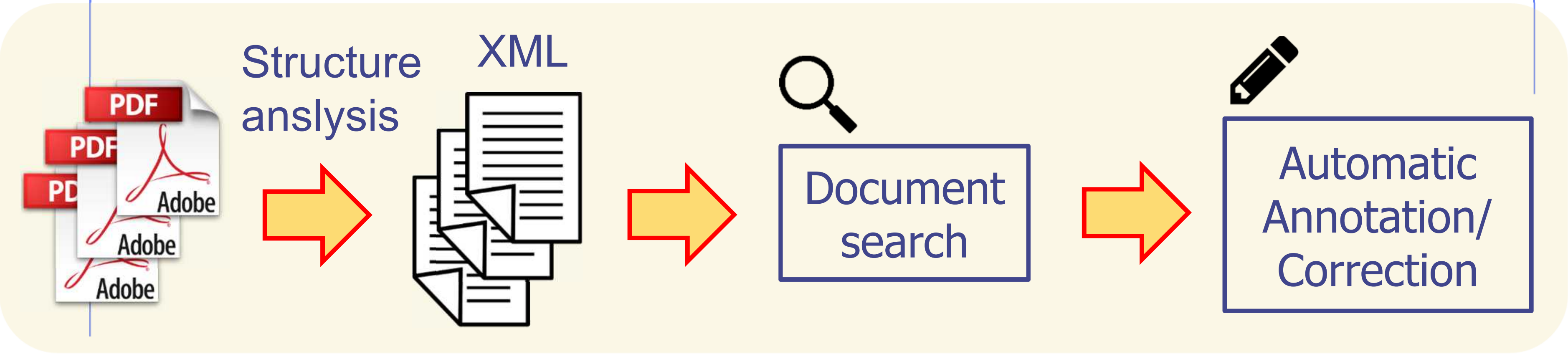
- Analysis
  - Document Analysis:
    - PDF, Tables, Graphs, Math formulas
  - Text Analysis: natural language analysis
  - Annotation tools
- Search
  - Aspect-based paper search
- Extraction
  - Concept / Relation / Event extraction
- Visualization
  - Citation relation / research trend
- Knowledge Base completion / Inference

# Recent achievements

- Analysis
  - PDF analysis tools
  - English Multi-word expression lexicon and MWE-aware text analysis tools [Kato et al, LREC-2018]
  - PDF / XML Annotation tools [Shindo et al, LREC-2018]
- Search
  - Aspect-based search and recommendation of papers [Kobayashi et al, JCDL-2018]
- Extraction
  - Relation extraction by distant supervision
  - Seed selection for distant supervision [Phi et al, ACL-2018]
- Visualization
  - Trend detection from citation network [Asatani et al, PLOS one 2018]
- Knowledge Base completion / Inference
  - Symmetric/Asymmetric relation acquisition [Manabe et al, AAAI-2018]

# Document Analysis Tools

## Overall systems and their relation



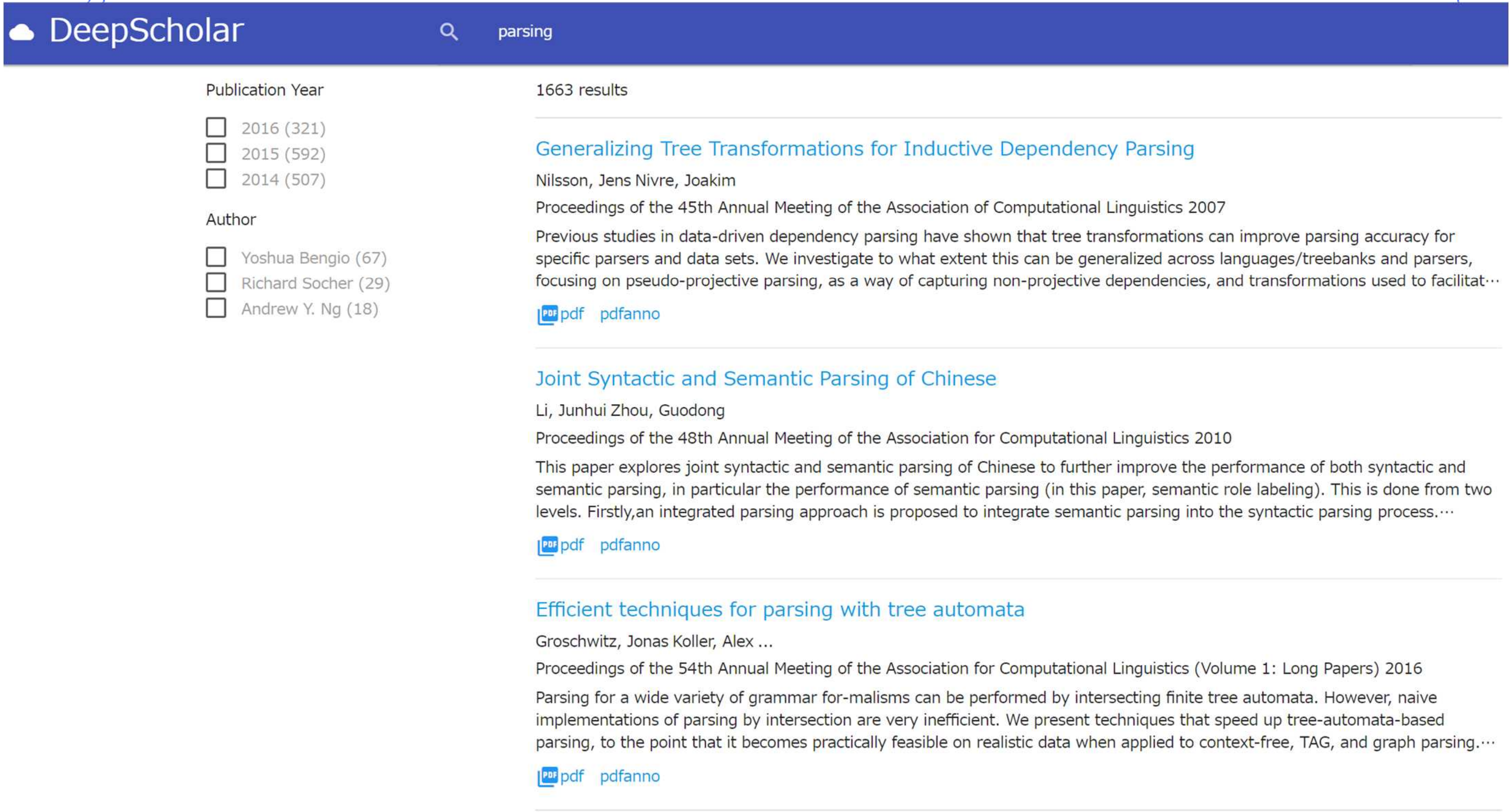
- PDFExtract
- PDF2XML
- Math formula analyzer
- In-line math expression analysis

- DeepScholar
- SideNoter
- Citation analysis
- DeepCRF(NER)
- Relation Extraction

- PDFAnno
- XMLAnno

# XML search engine

DeepScolar



Publication Year

□ 2016 (321)
□ 2015 (592)
□ 2014 (507)

Author

□ Yoshua Bengio (67)
□ Richard Socher (29)
□ Andrew Y. Ng (18)

1663 results

**Generalizing Tree Transformations for Inductive Dependency Parsing**

Nilsson, Jens Nivre, Joakim

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics 2007

Previous studies in data-driven dependency parsing have shown that tree transformations can improve parsing accuracy for specific parsers and data sets. We investigate to what extent this can be generalized across languages/treebanks and parsers, focusing on pseudo-projective parsing, as a way of capturing non-projective dependencies, and transformations used to facilitat…

pdf pdfanno

**Joint Syntactic and Semantic Parsing of Chinese**

Li, Junhui Zhou, Guodong

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics 2010

This paper explores joint syntactic and semantic parsing of Chinese to further improve the performance of both syntactic and semantic parsing, in particular the performance of semantic parsing (in this paper, semantic role labeling). This is done from two levels. Firstly,an integrated parsing approach is proposed to integrate semantic parsing into the syntactic parsing process.…

pdf pdfanno

**Efficient techniques for parsing with tree automata**

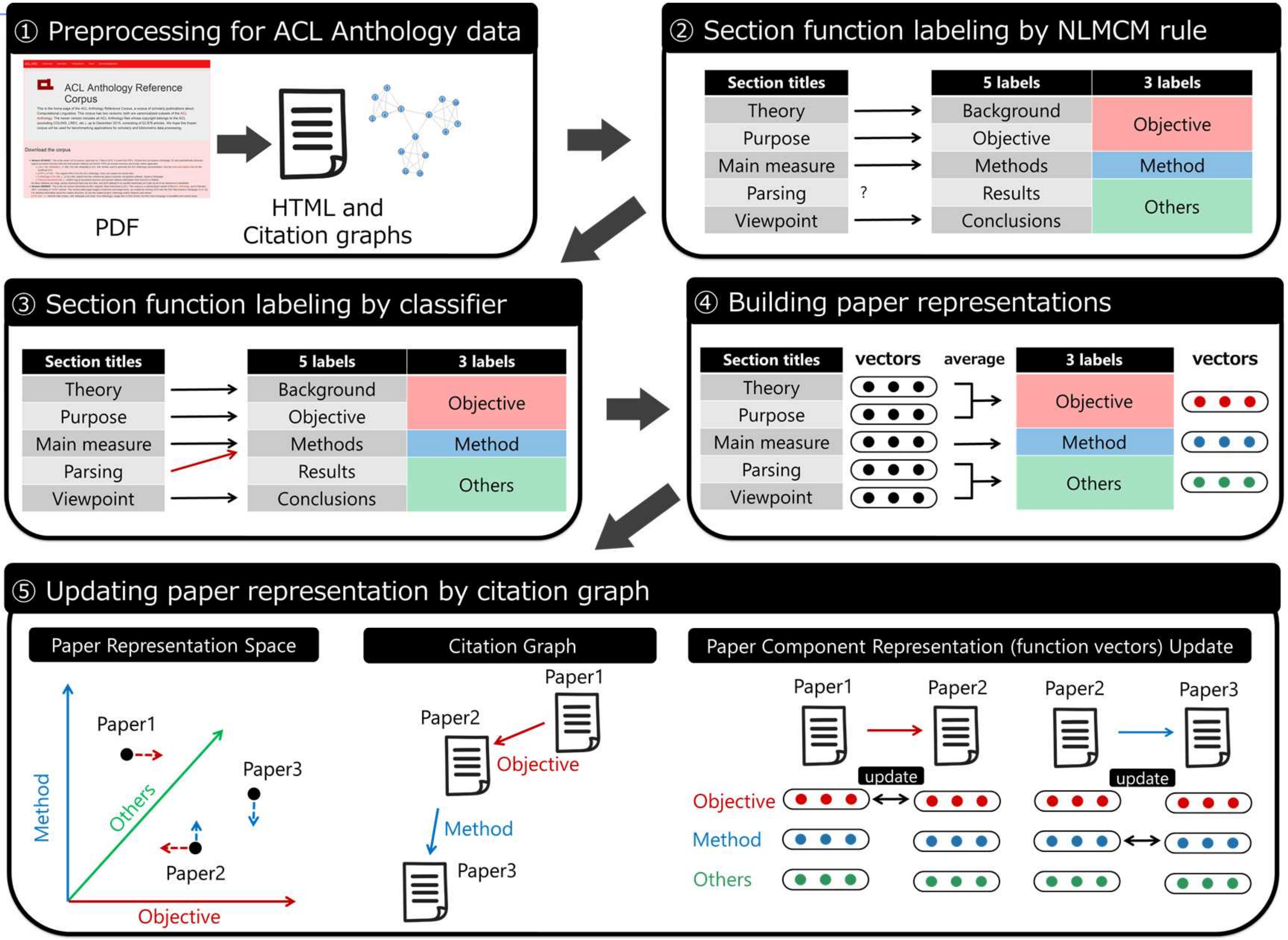
Groschwitz, Jonas Koller, Alex ...

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2016

Parsing for a wide variety of grammar for-malisms can be performed by intersecting finite tree automata. However, naive implementations of parsing by intersection are very inefficient. We present techniques that speed up tree-automata-based parsing, to the point that it becomes practically feasible on realistic data when applied to context-free, TAG, and graph parsing.…

pdf pdfanno

# Aspect-based similarity learning for paper retrieval [Kobayashi, Shimbo, Matsumoto 2018]



① Preprocessing for ACL Anthology data

② Section function labeling by NLMCM rule

| Section titles | | 5 labels | 3 labels |
|---|---|---|---|
| Theory | → | Background | Objective |
| Purpose | → | Objective | |
| Main measure | → | Methods | Method |
| Parsing | ? | Results | Others |
| Viewpoint | → | Conclusions | |

③ Section function labeling by classifier

| Section titles | | 5 labels | 3 labels |
|---|---|---|---|
| Theory | → | Background | Objective |
| Purpose | → | Objective | |
| Main measure | → | Methods | Method |
| Parsing | → | Results | Others |
| Viewpoint | → | Conclusions | |

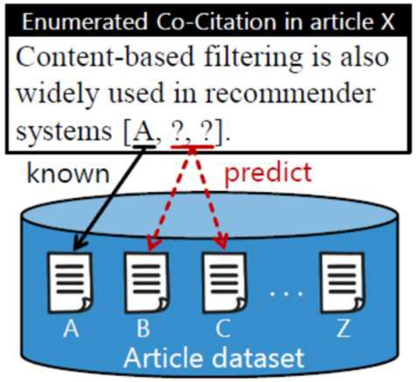
④ Building paper representations

| Section titles | vectors | average | 3 labels | vectors |
|---|---|---|---|---|
| Theory | ● ● ● | → | Objective | ● ● ● |
| Purpose | ● ● ● | | | |
| Main measure | ● ● ● | → | Method | ● ● ● |
| Parsing | ● ● ● | → | Others | ● ● ● |
| Viewpoint | ● ● ● | | | |

⑤ Updating paper representation by citation graph

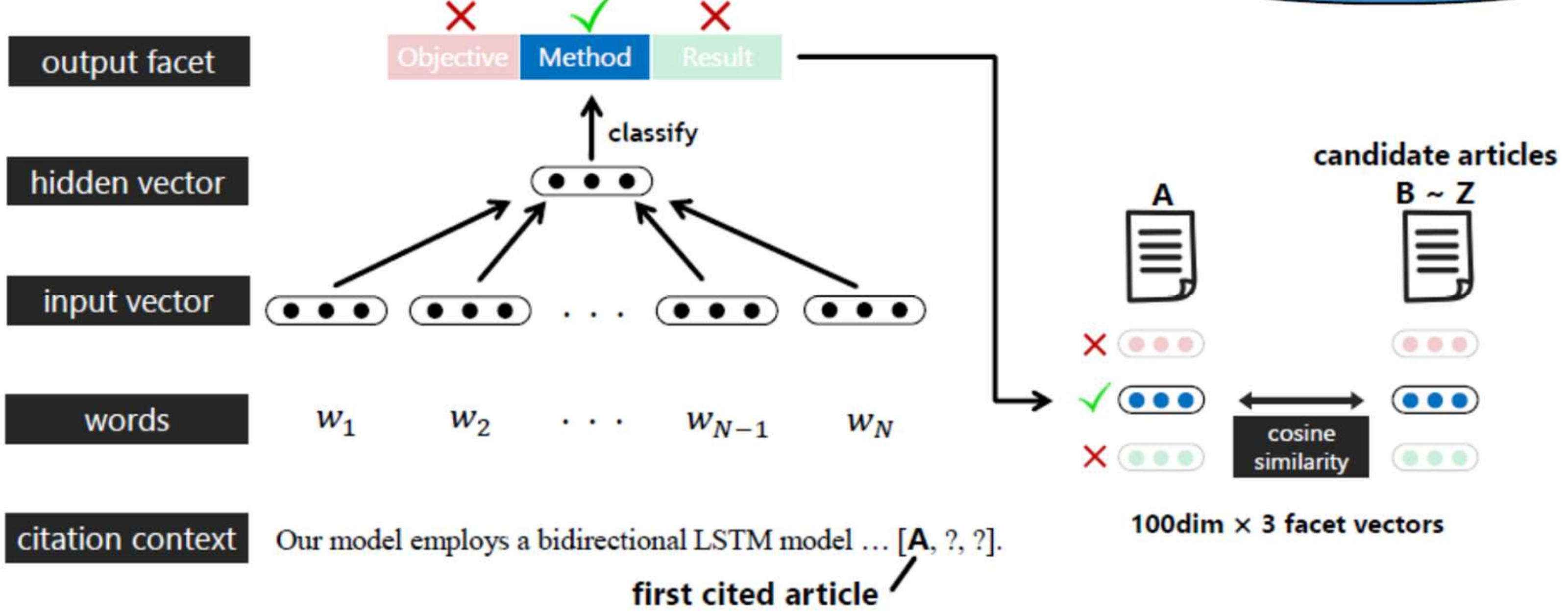
# Aspect-based similarity learning for paper retrieval [Kobayashi, Shimbo, Matsumoto 2018]

- Classification of citation contexts into aspects: → Objective, Method, Result
- Make recommendation of papers similar to the cited paper sharing the aspect



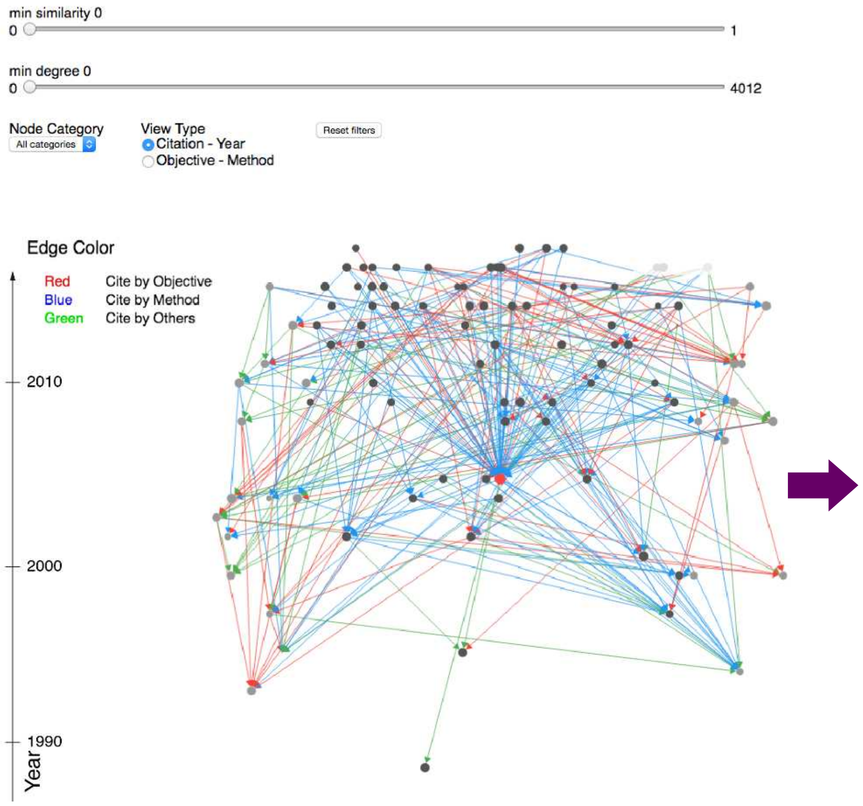
Enumerated Co-Citation in article X

Content-based filtering is also widely used in recommender systems [A, ?, ?].



Our model employs a bidirectional LSTM model … [A, ?, ?].

# Visualization of similar papers



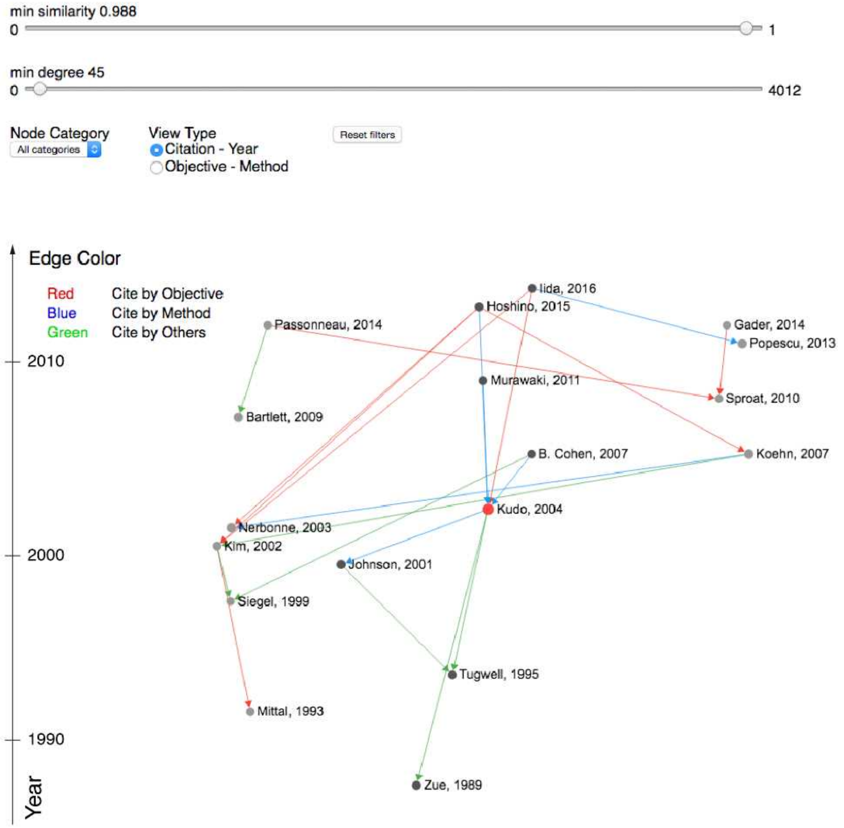
min similarity 0
0 ○ 1

min degree 0
0 ○ 4012

Node Category: All categories

View Type
● Citation - Year
○ Objective - Method

Reset filters



min similarity 0.988
0 ○ 1

min degree 45
0 ○ 4012

Node Category: All categories

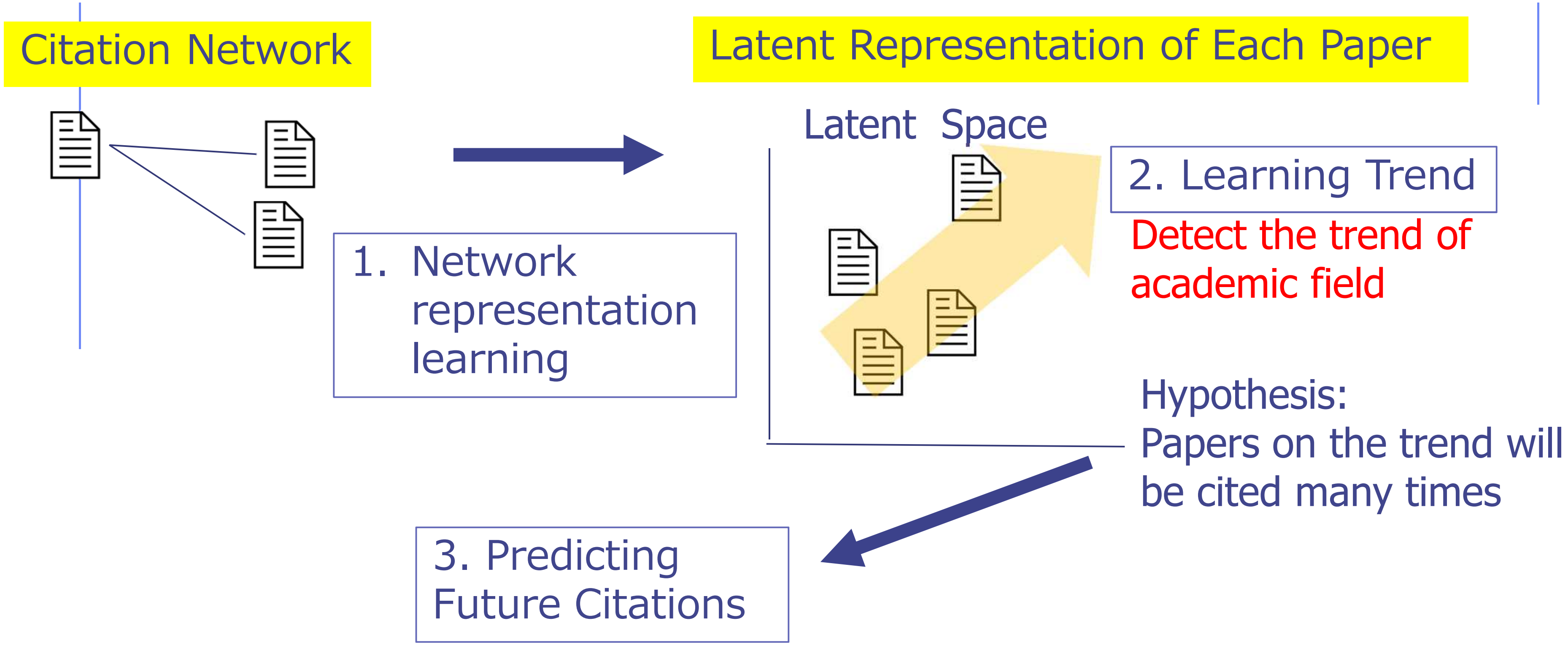
View Type
● Citation - Year
○ Objective - Method

Reset filters

User can control density of graphs by setting similarity threshold

# Trend Detection / Prediction

Detecting trend of academic fields using network representation learning [Asatani et al 2018]



Citation Network

Latent Representation of Each Paper

Latent Space

1. Network representation learning

2. Learning Trend

Detect the trend of academic field

Hypothesis:
Papers on the trend will be cited many times

3. Predicting Future Citations

Predict the number of citation better than PageRank.. etc

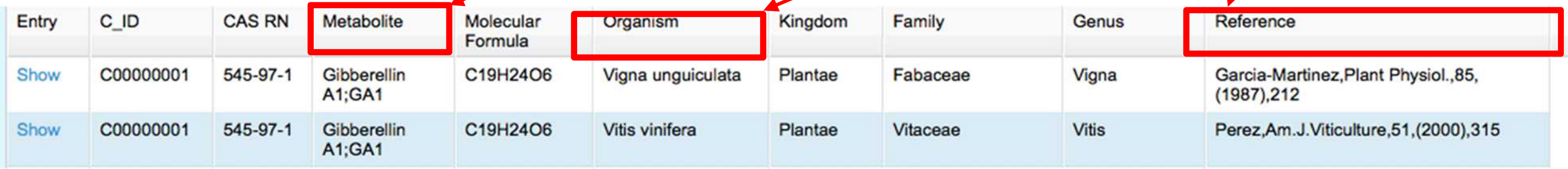
# Relation Extraction (Knowledge Acquisition)

- Knowledge Base Completion
  - KNApSAcK Database (NAIST)
    - Databese of Metabolite-Plant Species Relationship
  - KEGG Pathway (© Kyoto University)
    - Collection of manually drawn pathway maps representing knowledge on the molecular interaction, reaction and relation networks (substrate-enzyme-product)
  - Property extraction of Thermoelectric materials (NAIST)
    - Properties: electric conductivity, thermal conductivity, Seebeck coefficient, etc.

# Knowledge Base Completion by distant supervision

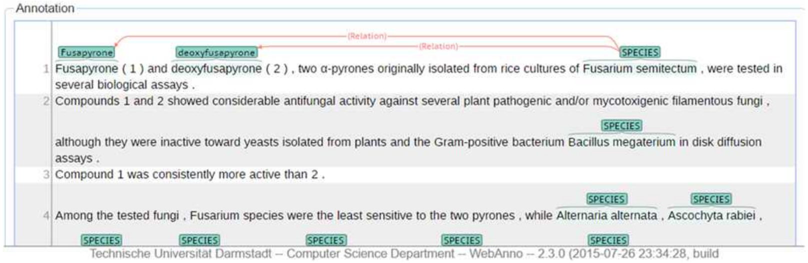
Existing Knowledge Base (KNApSAcK DB)

Metabolite Species Reference



| Entry | C_ID | CAS RN | Metabolite | Molecular Formula | Organism | Kingdom | Family | Genus | Reference |
|---|---|---|---|---|---|---|---|---|---|
| Show | C00000001 | 545-97-1 | Gibberellin A1;GA1 | C19H24O6 | Vigna unguiculata | Plantae | Fabaceae | Vigna | Garcia-Martinez,Plant Physiol.,85,(1987),212 |
| Show | C00000001 | 545-97-1 | Gibberellin A1;GA1 | C19H24O6 | Vitis vinifera | Plantae | Vitaceae | Vitis | Perez,Am.J.Viticulture,51,(2000),315 |

Automatic Annotation of Terms and Relations



Annotation

1 Fusapyrone ( 1 ) and deoxyfusapyrone ( 2 ) , two α-pyrones originally isolated from rice cultures of Fusarium semitectum , were tested in several biological assays .

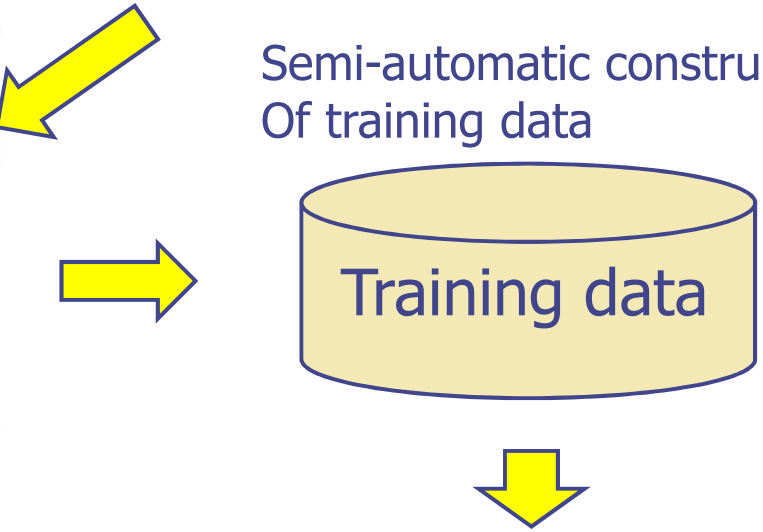
2 Compounds 1 and 2 showed considerable antifungal activity against several plant pathogenic and/or mycotoxigenic filamentous fungi , although they were inactive toward yeasts isolated from plants and the Gram-positive bacterium Bacillus megaterium in disk diffusion assays .

3 Compound 1 was consistently more active than 2 .

4 Among the tested fungi , Fusarium species were the least sensitive to the two pyrones , while Alternaria alternata , Ascochyta rabiei ,

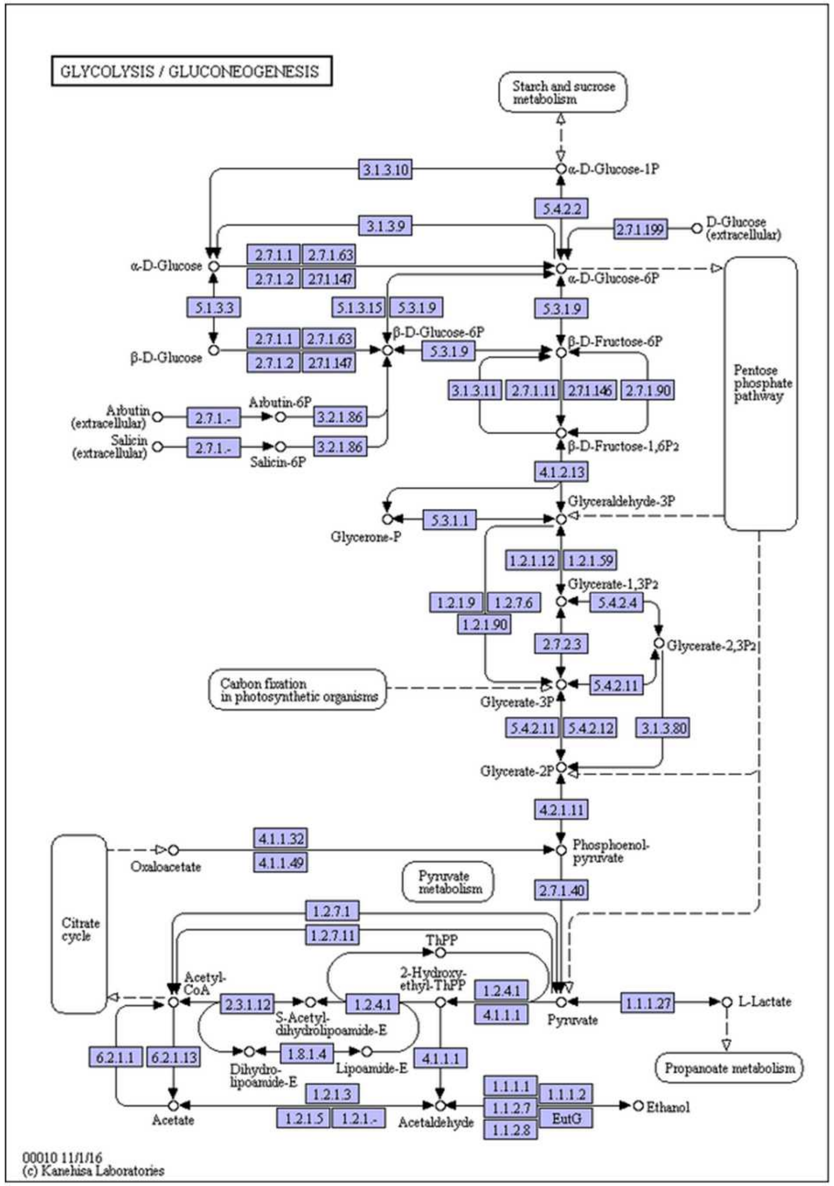
Technische Universität Darmstadt -- Computer Science Department -- WebAnno -- 2.3.0 (2015-07-26 23:34:28, build

Semi-automatic construction Of training data



Classifier / Sequence Labeler

# KEGG Pathway Map Generation

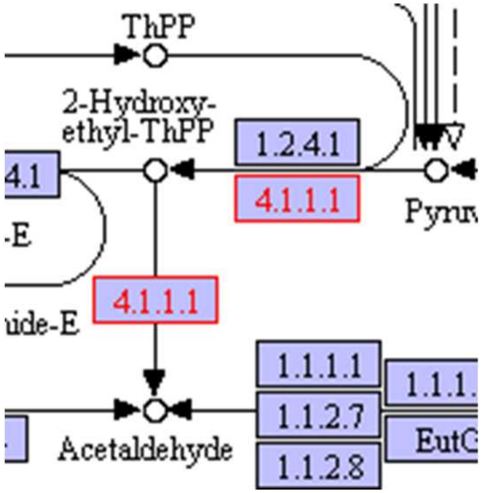


## KEGG: Kyoto Encyclopedia of Genes and Genomes

▭ : Gene products

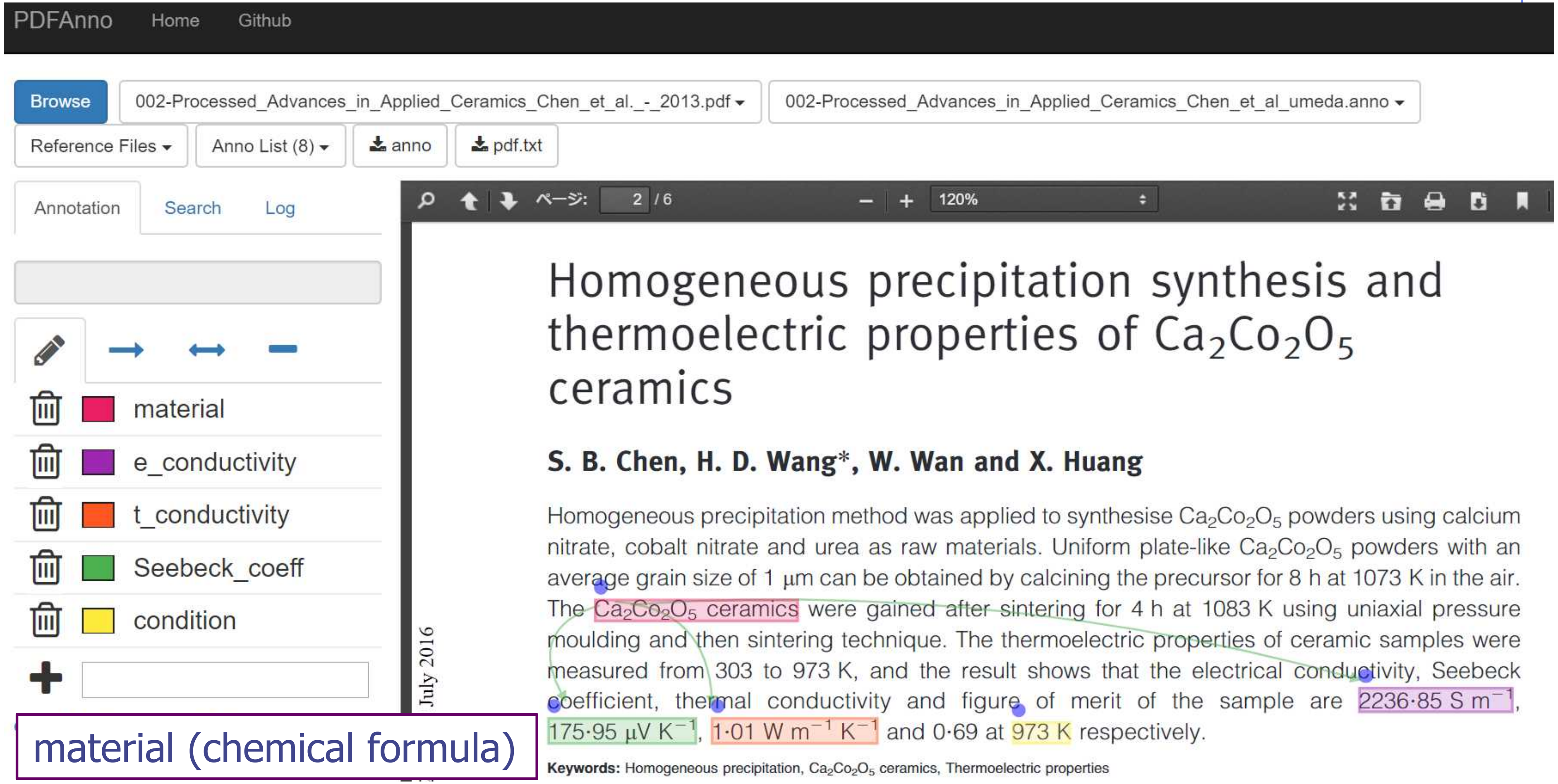
○ : Chemical compounds

- Relation (protein network)
- Reaction (chemical network)



# Property extraction of Thermoelectric materials

(PDFAnno: annotation / browsing tool [Shindo, Matsumoto 2018])


PDFAnno
Home
Github
Browse
002-Processed_Advances_in_Applied_Ceramics_Chen_et_al._-_2013.pdf
002-Processed_Advances_in_Applied_Ceramics_Chen_et_al_umeda.anno
Reference Files
Anno List (8)
anno
pdf.txt
Annotation
Search
Log
ページ: 2 / 6
120%
July 2016

🟥 material
🟪 e_conductivity
🟧 t_conductivity
🟩 Seebeck_coeff
🟨 condition

## Homogeneous precipitation synthesis and thermoelectric properties of $Ca_2Co_2O_5$ ceramics

**S. B. Chen, H. D. Wang*, W. Wan and X. Huang**

Homogeneous precipitation method was applied to synthesise $Ca_2Co_2O_5$ powders using calcium nitrate, cobalt nitrate and urea as raw materials. Uniform plate-like $Ca_2Co_2O_5$ powders with an average grain size of 1 µm can be obtained by calcining the precursor for 8 h at 1073 K in the air. The $Ca_2Co_2O_5$ ceramics were gained after sintering for 4 h at 1083 K using uniaxial pressure moulding and then sintering technique. The thermoelectric properties of ceramic samples were measured from 303 to 973 K, and the result shows that the electrical conductivity, Seebeck coefficient, thermal conductivity and figure of merit of the sample are 2236·85 S $m^{-1}$, 175·95 µV $K^{-1}$, 1·01 W $m^{-1}$ $K^{-1}$ and 0·69 at 973 K respectively.

**Keywords:** Homogeneous precipitation, $Ca_2Co_2O_5$ ceramics, Thermoelectric properties

material (chemical formula)

properties

electrical conductivity (σ)

thermal conductivity (κ)

Seebeck coefficient (S)

condition

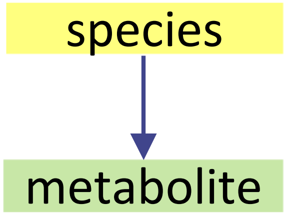
temperature (T)

# Some Ongoing Projects

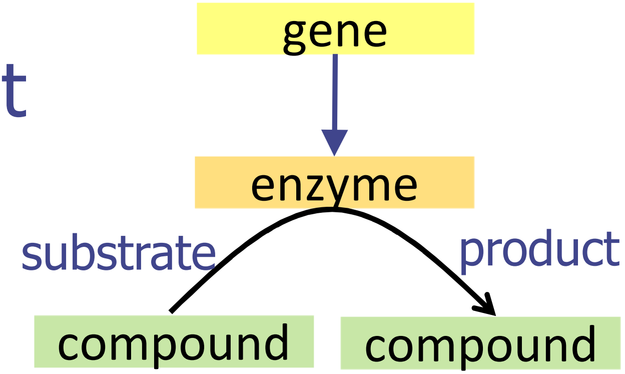
- Knowledge Structure Acquisition from papers in the same domain
  - Relation (pairwise, multi-items)
  - Experimental settings, Synthesis processes
- Review Matrix Generation
- Acquisition of pros and cons of methodologies

# Knowledge Structure Acquisition

- KNApSAcK DB: relation between species and metabolies
- KEGG: relation between gene, enzyme, substrate and product
- Thermoelectric materials


species

metabolite


gene

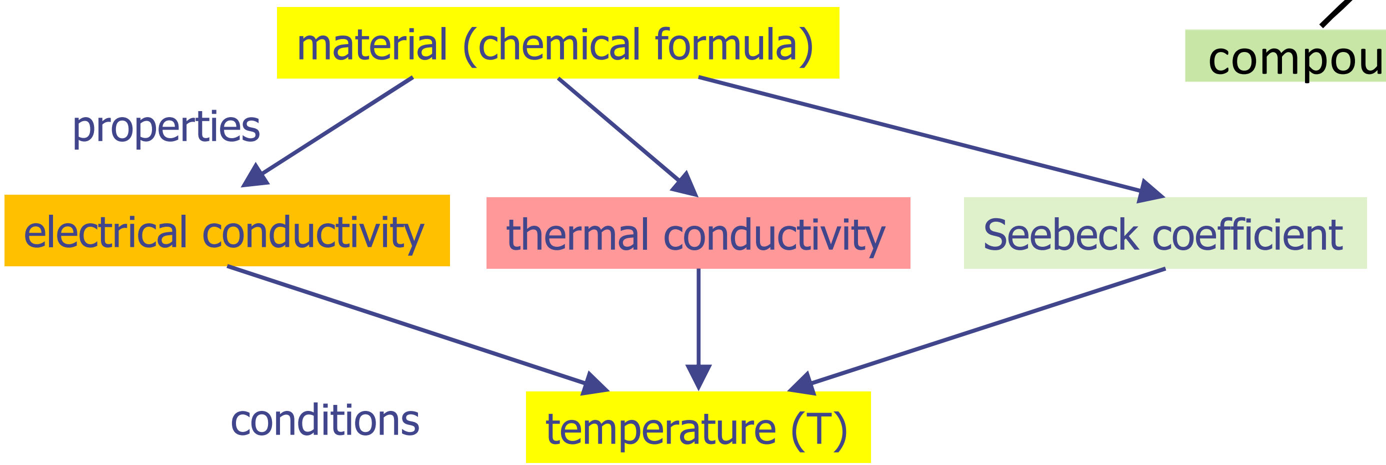
enzyme

substrate

product

compound

compound


material (chemical formula)

properties

electrical conductivity

thermal conductivity

Seebeck coefficient

conditions

temperature (T)

# Document and Text Analysis

- Document Analysis
  - PDF analysis
    - XML conversion, Table / Graph / Math formula analysis
  - Citation analysis / Document similarity
- Text Analysis
  - Base NLP analysis tools: POS tagging, parsing, NE recognition, Relation extraction, Predicate-argument structure analysis
  - Complex sentence structure analysis
- Applications
  - Document retrieval / Visualization
  - Knowledge Base completion
  - Comparative study of scientific papers

# Summary

- Scientific Document Analysis
  - PDF, Tables, Graphs, Math formula Analysis
  - Text Analysis: natural language analysis
  - Annotation tools
- Search
  - Aspect-based paper search
- Extraction
  - Concept / Relation / Event extraction
  - Common Structure Acquisition
- Visualization
  - Citation relation / research trend
- Knowledge Base completion