# 特定领域知识图谱构建初探

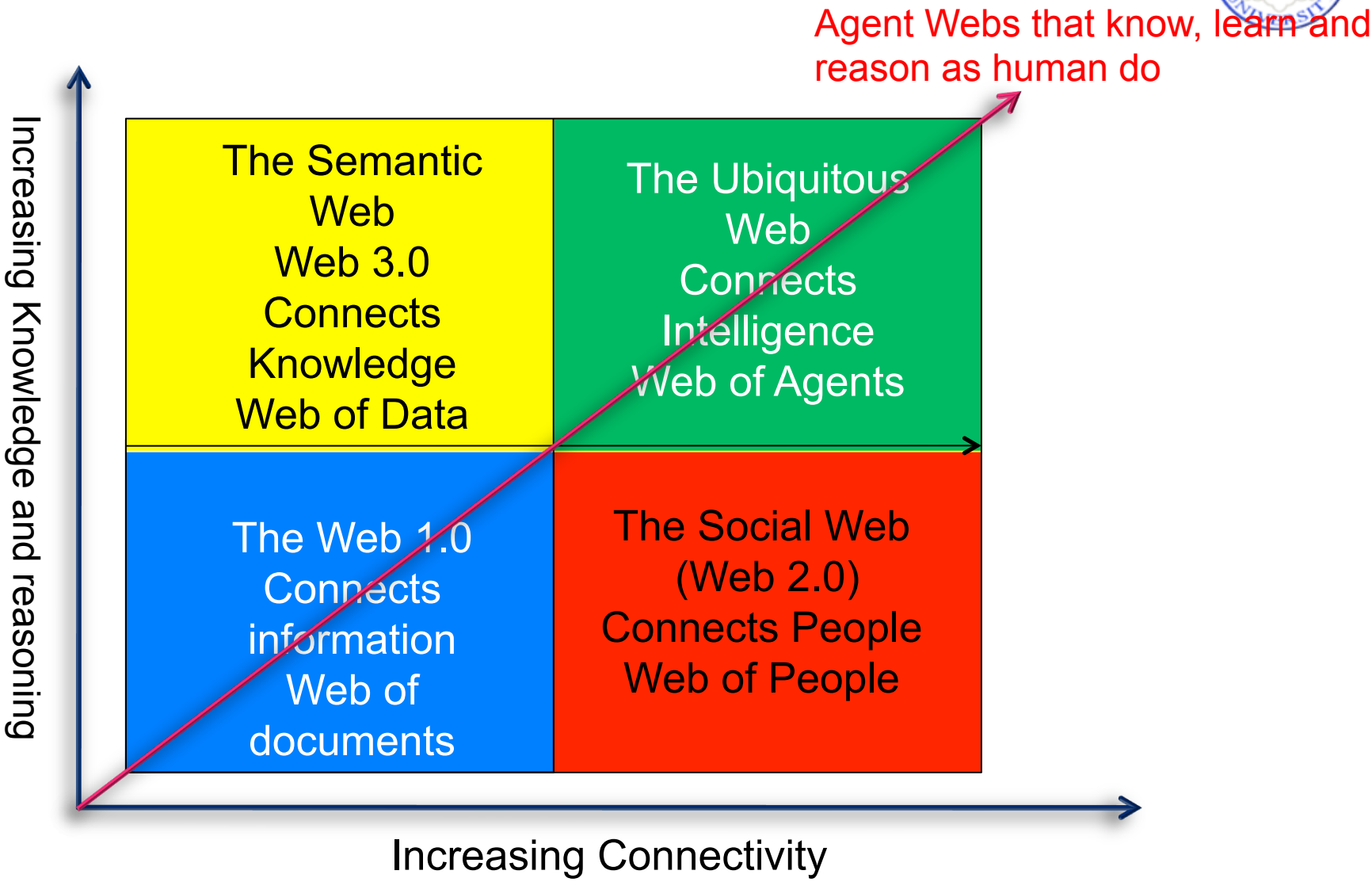
李涓子

清华大学计算机系知识工程研究室

# Outline

- Knowledge graph and technologies
- Big scholar knowledge base - Aminer II
- Knowledge graph building over enterprise data
- Conclusion

# Vision of future Web

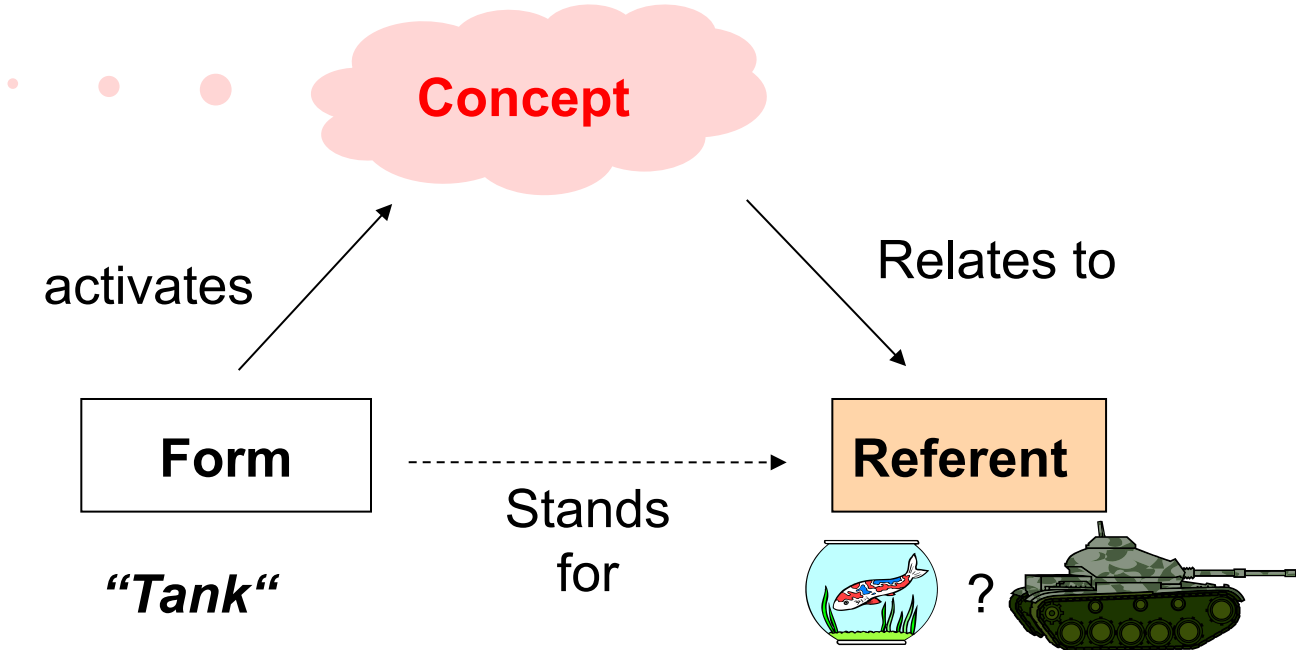


# Bring structure to the meaningful content of Web pages


The Semantic Web. Tim Berners-Lee, James Hendler, and Ora Lassila. Scientific American, 2001.

# Philosophy of ontology

## Concept triangle




**[Ogden, Richards, 1923]**

*Ontology is the philosophical study of the nature of **being**, **becoming**, **existence**, or **reality**, as well as the **basic categories** of being and their relations.*

--- Wikipedia

# Some knowledge graphs

DBpedia

250 concepts
4M instances
6000 properties
500 Triples

yago
select knowledge

350K Cs
10M Is
100 Ps
120M Ts

搜狗搜索

BabelNet 1.1.1
A very large multilingual ontology

50M Ss
50+Ls
262M Ts

Baidu 百度

Freebase

15K Cs
40M Is
4000 Ps
1BTs
Google KB Core

WolframAlpha

NELL

XLORE

850K Cs
8M Is
70K Ps

OpenIE
(Reverb, OLLIE)

Google KG

15K Cs
600M Is
20B Ts

WordNet
7 Europe Ls
Cross lingual links

EuroWordNet

# Our Knowledge graph definition

- *C – concepts*
  - A group of objects with same properties
  - cars, students, professors
- *I - instances*
  - A object which belongs to a concept
  - **Peter** is a student
- *T – ISA*
  - **subConceptOf, instanceOf**
- *P – properties*
  - char本体中用于描述实例信息的**其他语义关系**
  - 如：**instance-attribute-value** (AVP)

**Taxonomy Knowledge**

**Factual knowledge**

# Knowledge graph technologies

- **Manually KG building: Wordnet, Cyc, Hownet**
- **Taxonomy knowledge learning**
  - Learning from Wikipedia
  - Learning beyond Wikipedia
- **Factual knowledge learning**
  - Learning from Wikipedia
  - Learning beyond Wikipedia

# Learning taxonomy knowledge from Wikipedia

## Category system in Wikipedia

- Category system in Wikipedia as a conceptual network

  PHILOSOPHY and BELIEF (deals-with?)

  PHILOSOPHY and HUMANITIES (isa)

  PHILOSOPHY and SCIENCE (isa)

Advantages:

- widely recognized – concepts in human minds
- Large scale - over millions of concepts and ten millions of instances
- Large coverage

Problems:

- noise – categories for different purposes
- inconsistence - not well formally define

# Learning taxonomy knowledge from Wikipedia

- Using linguistic features of isa relationship
  - syntactic parsing:

    head matching/modifier matching/Singular/plural forms
  - Lexico-patterns:
- Using structure of wikipedia

Deriving a Large Scale Taxonc
al. AAAI 07.

- Using external high quality is

  wordnet, Hownet, Cilin

  YAGO(WWW2007)
- isa relation validation using c

  Xlore (AAAI2014)

1. *NP2, ?* (such as|like|, especially) *NP* NP1*
   *a stimulant such as caffeine*
2. such *NP2* as *NP* NP1*
   *such stimulants as caffeine*
3. *NP1 NP** (and|or|,like) other *NP2*
   *caffeine and other stimulants*
4. *NP1*, one of *det_pl NP2*
   *caffeine, one of the stimulants*
5. *NP1*, *det_sg NP2 rel_pron*
   *caffeine, a stimulant which*
6. *NP2* like *NP* NP1*
   *stimulants like caffeine*

# Learning taxonomy knowledge beyond Wikipedia

- **Using Web sources**

  Root concepts, search engine

  - Hearst patterns
  - Bootstrapping
  - Taxonomy induction (structural learni

  domain specific taxonomy building

  EMNLP2010, ACL 2014

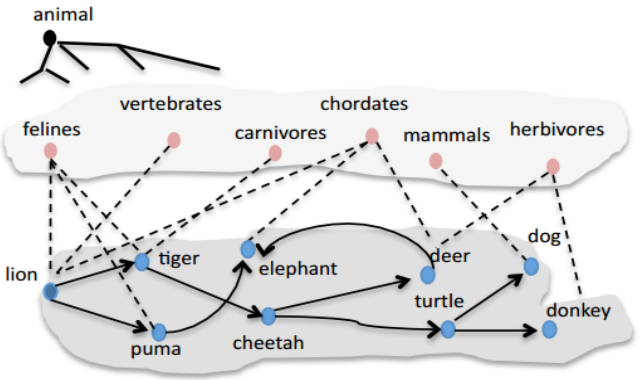


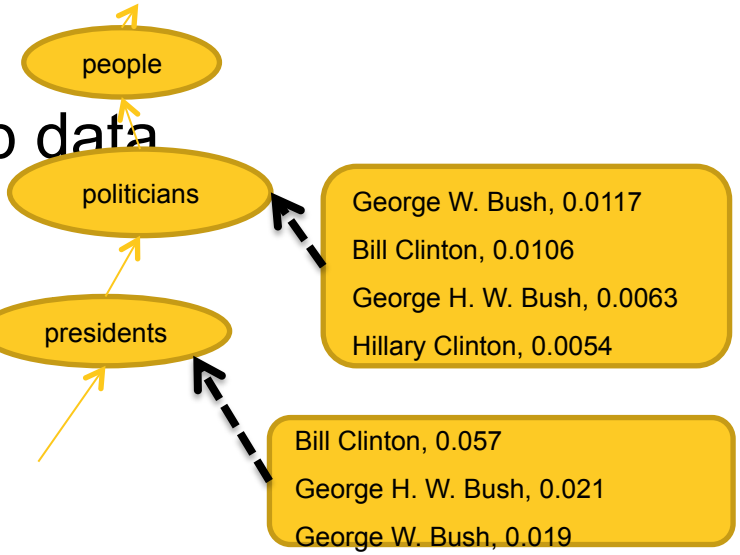

Figure 1: *Taxonomy Induction from Scratch.*

- **Large scale taxonomy building**

  - Automatically generated from Web data
  - 1.6 billion web pages
  - Rich hierarchy of millions of conc
  - Probabilistic knowledge base

  SIGMOD2012

  - Probase: **2,653,872 concepts** **20,757,545 Isa**



# Factual knowledge learning

| | Supervised | Semi-supervised | Unsupervised |
|---|---|---|---|
| From Wikipedia | Sematic annotation | Semantify Wikipdia-Kylin<br>Cross lingual IE-WikiCiKE | |
| Beyond Wikipedia | | Distant supervision(Stanford)<br>Coupled Semi-Supervised Learning(NELL) | KnowItAll:<br>TextRuner<br>WOE |
| | | | |

# Automatic semantic annotation

- Rule learning based approach

  Automatically learn annotation rules from the training data

- Classification based approach

  Identify the boundary of tags in instances using classification models

- Sequential labeling based approach

Consider the dependencies between tags

- Constrained Hierarchical Conditional Random Fields
- And Others ….

# Learning factual knowledge beyond Wikipedia-Knowledge Vault

- Current large scale knowledge graph is still not enough

| Relation | % unknown in Freebase |
|---|---|
| Profession | 68% |
| Place of birth | 71% |
| Nationality | 75% |
| Education | 91% |
| Spouse | 92% |
| Parents | 94% |

| Name | # Entity types | # Entity instances | # Relation types | # Confident facts (relation instances) |
|---|---|---|---|---|
| *Knowledge Vault (KV)* | 1100 | 45M | 4469 | 271M |
| DeepDive [30] | 4 | 2.7M | 34 | 7M[a] |
| NELL [8] | 271 | 5.19M | 306 | 0.435M[b] |
| PROSPERA [28] | 11 | N/A | 14 | 0.1M |
| YAGO2 [18] | 350,000 | 9.8M | 100 | 4M[c] |
| Freebase [4] | 1,500 | 40M | 35,000 | 637M[d] |
| Knowledge Graph (KG) | 1,500 | 570M | 35,000 | 18,000M[e] |

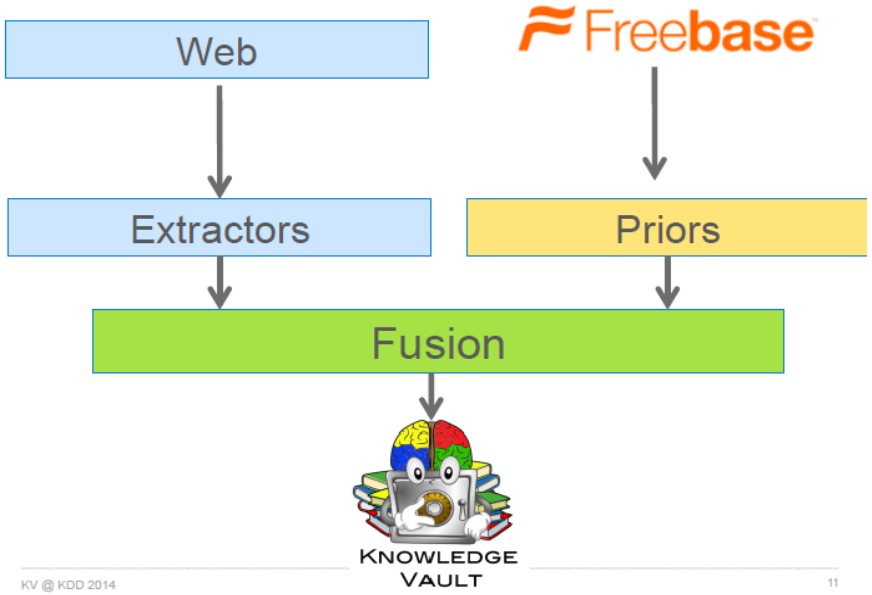
Table 1: Comparison of knowledge bases. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Confident facts means with a probability of being true at or above 0.9.

# Learning factual knowledge beyond Wikipedia-Knowledge Vault

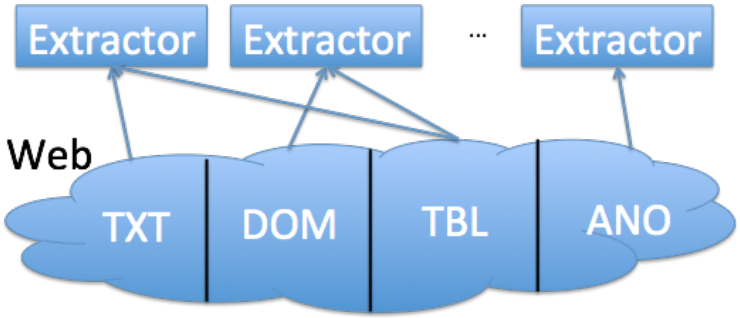
## Motivation

- the new approach should automatically leverage already-cataloged knowledge to build prior models of fact correctness

## Framework




TXT: Distant supervision
DOM: DOM tree structure features
TBL： Table information
ANO: annotated tags in htmls

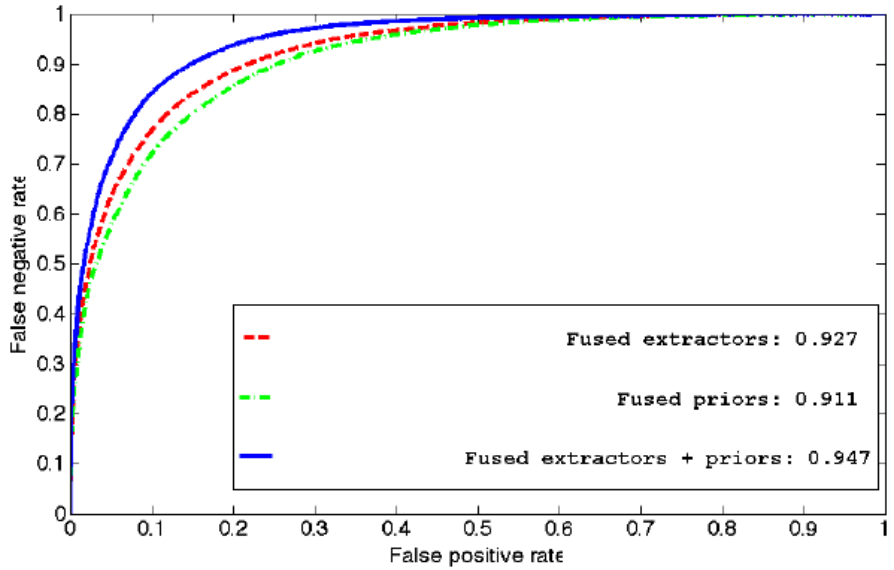



Priors: Path ranking algorithm
Priors: Neural network method

# Learning factual knowledge beyond Wikipedia-Knowledge Vault

| System | # | # > 0.7 | # > 0.9 | Frac. >0.9 | AUC |
|---|---|---|---|---|---|
| TBL | 9.4M | 3.8M | 0.59M | 0.06 | 0.856 |
| ANO | 140M | 2.4M | 0.25M | 0.002 | 0.920 |
| TXT | 330M | 20M | 7.1M | 0.02 | 0.867 |
| DOM | 1200M | 150M | 94M | 0.08 | 0.928 |
| FUSED-EX. | 1600M | 160M | 100M | 0.06 | 0.927 |

Table 2: Performance of different extraction systems.



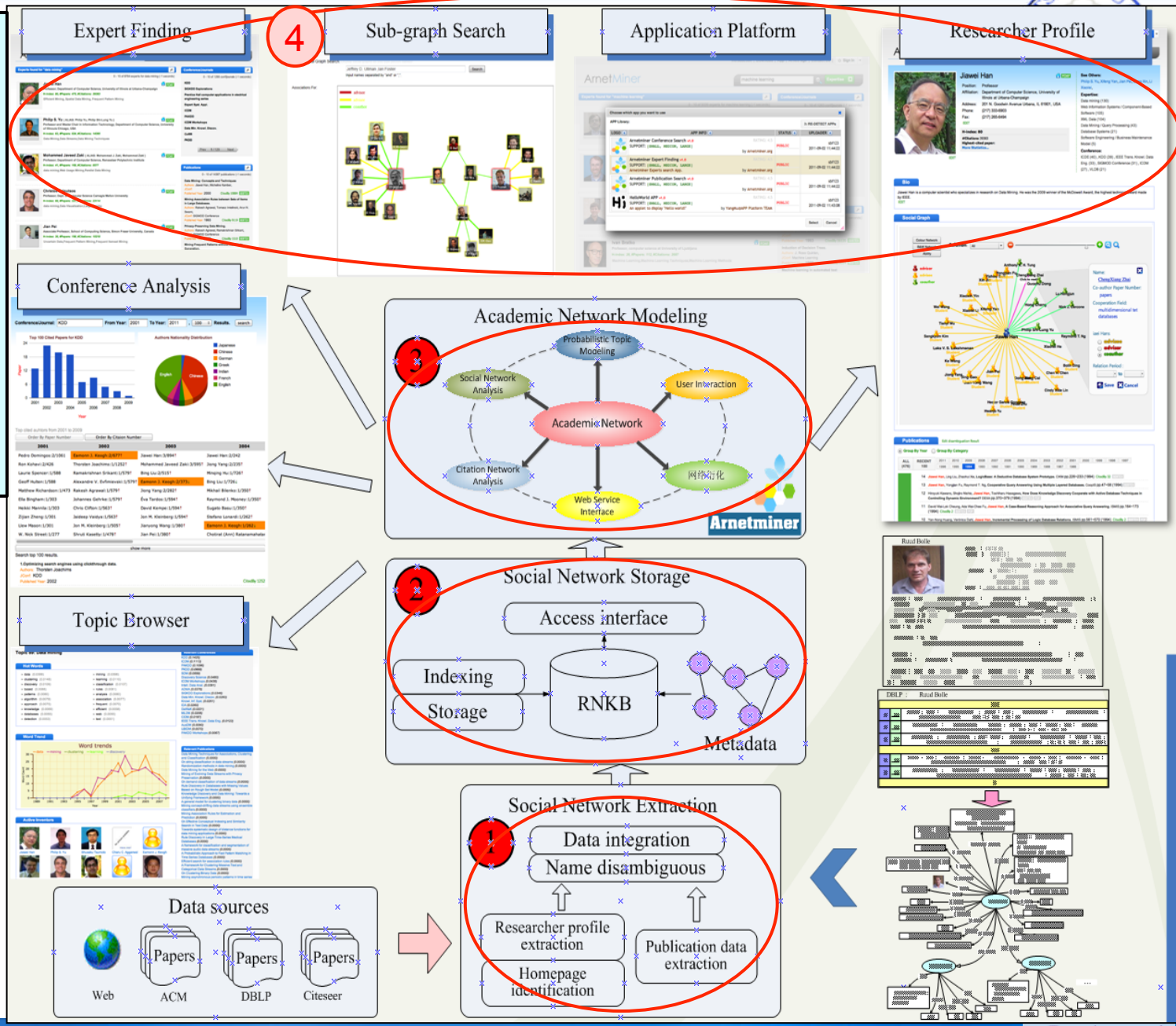
# Summary

- Various knowledge representation and learning methods
- What are the effective methods used for domain specific knowledge graph building?
- What are the proper representation for domain specific knowledge graph?

# Outline

- Knowledge graph and technologies
- Big scholar knowledge base - Aminer II
- Knowledge graph building over enterprise data
- Conclusion

- Researcher profile extraction
- Expert finding
- Social network search
- Topic browser
- Conference analysis
- ArnetApp platform



# Person Search


Upload

Jiawei Han (韩家炜)

Follow

Department of Computer Science, University of Illinois at Urbana-Champaign

Professor

(217) 333-6903

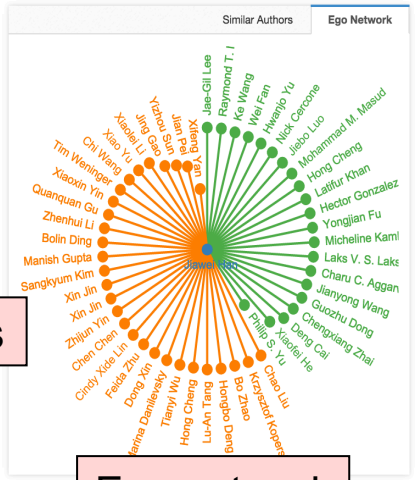
hanj@cs.uiuc.edu

http://www.cs.uiuc.edu/~hanj/

External Links

Update

Basic Info.




Ego network

Watch +

**Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies**

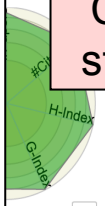
20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), New York 2014

Watch +

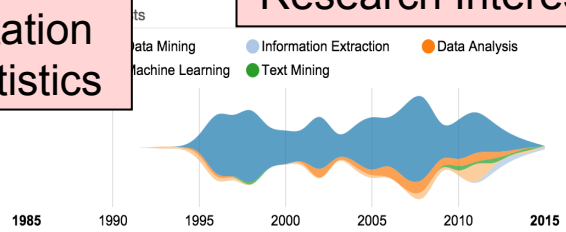
**Mining Massive RFID, Trajectory, and Traffic Data Sets**

14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas 2008

Citation statistics




Research Interests




About

Papers 790

Lectures 13

Patents 1

Add Paper | Remove Paper | By Year | By Citation

All (790) | Recent 20 | 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001 2000 1999 1998 1997 1996 1995 | 1994 1993 1992 1991 1990 1989 1988 1987 1986 1985

790

**Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions** Cited by 4

Wei Shen, Jianyong Wang, **Jiawei Han**

Knowledge and Data Engineering, IEEE Transactions (2015)

//dx.doi.org/10.1109/TKDE.2014.2327028

**Patterns.** Cited by 20

eng Yan, **Jiawei Han**

Mining (2014)

//dx.doi.org/10.1007/978-1-4419-6045-0_12

**Troubleshooting interactive complexity bugs in wireless sensor networks using data mining techniques** Cited by 5

Mohammad Maifi Hasan Khan, Hieu Khac Le, Hossein Ahmadi, Tarek F. Abdelzaher, **Jiawei Han**

ACM Transactions on Sensor Networks (TOSN) (2014)

Bibtex http://dx.doi.org/10.1145/2530290

787

1

**Systems and Methods for Detecting a Novel Data Class**

Mohammad Mehedy Masud, Latifur Rahman Khan, Bhavani Marienne Thuraisingham, Qing Chen, Jing Gao, Jiawei Han

Publication-date: 2012-03-01 Application-date: 2011-08-22

# Expert Search

Finding experts, for “data mining”

Demographics: gender, language, location, etc.

Knowledge about “data mining”

similar authors

AMiner | data mining | Home

Expert | Paper

Search Results for data mining

H-Index : >=60 (31) | 50-60 (19) | 40-50 (52) | 30-40 (107) | 20-30 (167) | 10-20 (403) | <10 (199)

Gender : Male (934) | Female (44)

Language : Chinese (290) | English (194) | Greek (37) | German (27) | French (23) | Japanese (20) | Korean (14) | Indian (12)

Location : USA (219) | China (141) | Taiwan (34) | Australia (33) | Canada (29) | Japan (25) | Germany (24) | Italy (20) | Hong Kong (20) | Singapore (20)

Relevance | H-Index | A-Index | Activity | Diversity | Rising Star | #Citation | #Paper

Similar

**Jiawei Han** (韩家炜) Follow

H-Index: 126 | #Paper: 790 | #Citation: 90481

Department of Computer Science, University of Illinois at Urbana-Champaign

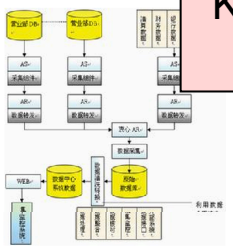
Professor

Data Mining | Information Extraction | Text Mining

Similar Authors:

2567 views

Data mining



Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

**Super Concepts:**

Data analysis | Data mining | Formal sciences | Applied sciences | Networks | Artificial intelligence

**Related Concepts:**

Data compression | Data visualization | Natural language processing | Data cleansing | Distributed computing | Informatization | Speech recognition | Business intelligence | Pattern recognition | Spatial database | Full text search | Metadata | Computer vision | ISAM | Biological neural network | Database | Grid computing | Database marketing | Parallel computing

必应 Jiawei Han Department of Computer Science University of Illinois at U

Web | Images | Videos | **Academic** | Dict | Knows | Maps | Explore

Time: Any Time | Since 2015 | Since 2014 | Since 2011 | Before 2011

Type: All Results | Paper Only

**Jiawei Han** - 微软学术

学者：Jiawei Han

机构：Department of Computer Science University of Illinois at Urbana Champaign

领域：Data Mining · Information Extraction

发表数：794 引用数：90537

Similar

**Xindong Wu** Follow

H-Index: 45 | #Paper: 331 | #Citation: 9644

Department of Computer Science, University of Vermont

Professor

Machine Learning | Information Extraction | Bayesian Networks | Data Mining | Supervised Learning

35 views

数据挖掘

数据挖掘(Data Mining)是通过分析每个数据，从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示3个步骤。数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异群组分析和演变分析等。

**上位词:**

资料分析 | 计算机科学基础理论 | 决策支持系统 | 信息管理术语 | 数据挖掘 | 形式科学

# Conference Ranking

AMiner | Whatever comes to your mind | Home

Computer Science

- All
- High Performance Computing
- Computer Network
- Net and Information Security

| Rank | Conference (Full Name) | Short Name | Impact Factor |
|---|---|---|---|
| 1 | Science | | 162.00 |
| 2 | Nucleic Acids Research | NAR | 128.00 |
| 3 | IEEE Conference on Computer Vision and Pattern Recognition | CVPR | 112.00 |
| 4 | IEEE Transactions on Pattern Analysis and Machine Intelligence | TPAMI | 101.00 |
| 5 | NeuroImage | NeuroImage | 99.00 |
| 6 | IEEE Transactions on Industrial Electronics | TIE | 80.00 |
| 7 | Bioinformatics | Bioinformatics | 79.00 |
| | munications Magazine | ICM | 74.00 |
| | sactions on Signal Processing | TSP | 73.00 |
| | erence on Human Factors in Computing Systems | CHI | 71.00 |
| | tology ePrint Archive | | 71.00 |
| | sactions on Image Processing | TIP | 69.00 |
| | a | Automatica | 69.00 |
| | national Conference on Computer Communications | INFOCOM | 69.00 |
| | cations of the ACM | Commun. ACM | 69.00 |
| | ns on Automatic Control | TAC | 67.00 |
| | al World Wide Web Conferences | WWW | 66.00 |
| | gs of the IEEE | Proc. IEEE | 64.00 |

## ACM Knowledge Discovery and Data Mining

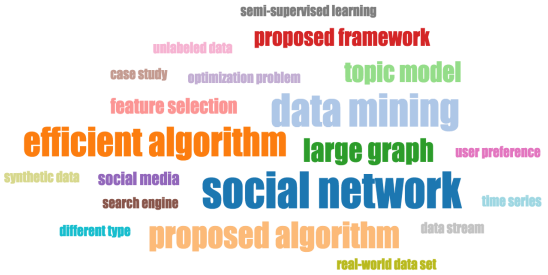
Search for Conference

Conference/Journal: ACM Knowledge Discovery and Data Mining

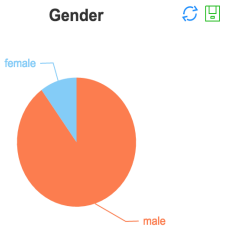
From Year: 2008

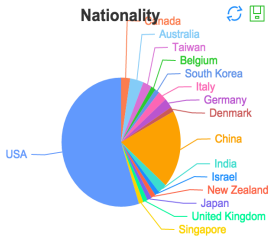
To Year: 2013

Submit




### Author Distribution



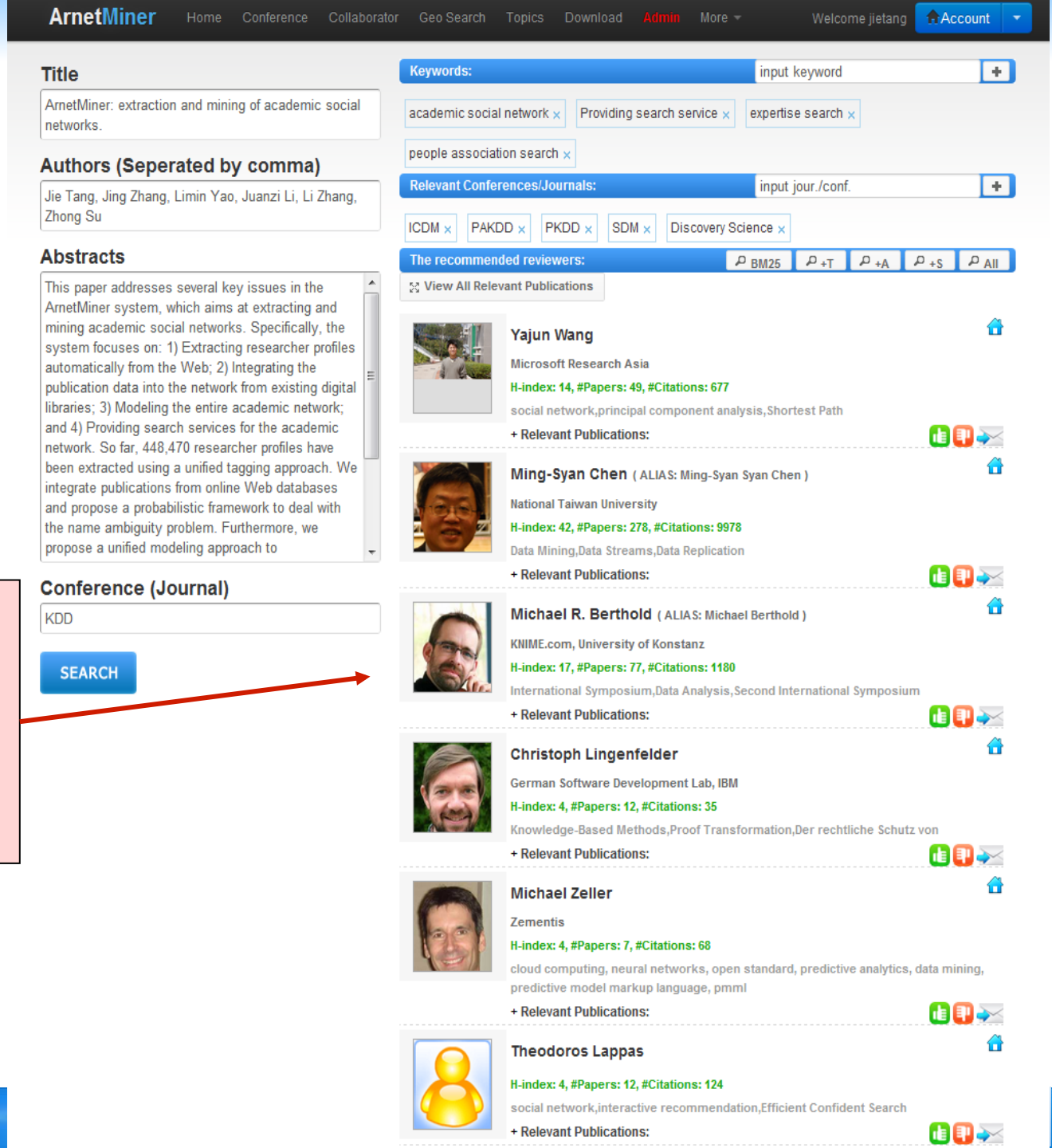


### Top cited authors

| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|---|
| Yehuda Koren:1/924 | Jure Leskovec:1/778↑ | Wei Chen:1/437↑ | Eunjoon Cho:1/541↑ | Jing Yuan:1/137↑ | Chris Thornton:1/64↑ |
| Jure Leskovec:1/474 | Wei Chen:1/578↑ | Yu Wang:1/168↑ | Dashun Wang:1/208↑ | Thanawin Rakthanmanon:1/108↑ | Bin Liu:1/33↑ |
| Jie Tang:1/440 | Jie Tang:1/384 | Hongning Wang:1/162↑ | Jing Yuan:1/171↑ | Alan Ritter:1/103↑ | Yu Zheng:1/33↑ |
| Yabo Xu:1/423 | Mohsen Jamali:1/312↑ | Deng Cai:1/135↑ | Rainer Gemulla:1/167↑ | Isabelle Stanton:1/86↑ | Hongzhi Yin:1/31↑ |
| Victor S. Sheng:1/420 | Justin Ma:1/259↑ | Maayan Roth:1/131↑ | Salvatore Scellato:1/164↑ | Ashton Anderson:1/78↑ | Arjun Mukherjee:1/31↑ |
| David Crandall:1/404 | Albert Bifet:1/249↑ | Liang Xiang:1/128↑ | Marco Pennacchiotti:1/106↑ | Ling-Yin Wei:1/75↑ | Charalampos Tsourakakis:1/27↑ |
| Aris Anagnostopoulos:1/392 | Theodoros Lappas:2/235↑ | Arik Friedman:1/125↑ | Noman Mohammed:1/93↑ | Rui Li:1/75↑ | Quan Yuan:1/26↑ |
| Huanhuan Cao:1/313 | Anna Monreale:1/229↑ | Min-Ling Zhang:1/113↑ | Mao Ye:1/90↑ | Jie Tang:2/68↑ | Reza Zafarani:1/24↑ |
| David J. Crandall:1/218 | Sayali Kulkarni:1/228↑ | Yong Ge:1/104↑ | Wei Liu:1/89↑ | Jiliang Tang:2/67↑ | Madhav Jha:1/23↑ |
| Ian Porteous:1/216 | Prem Melville:1/207↑ | Michael Jahrer:1/103↑ | Robson Leonardo Ferreira Cordeiro:1/74↑ | Xiwang Yang:1/65↑ | Wook-Shin Han:1/22↑ |

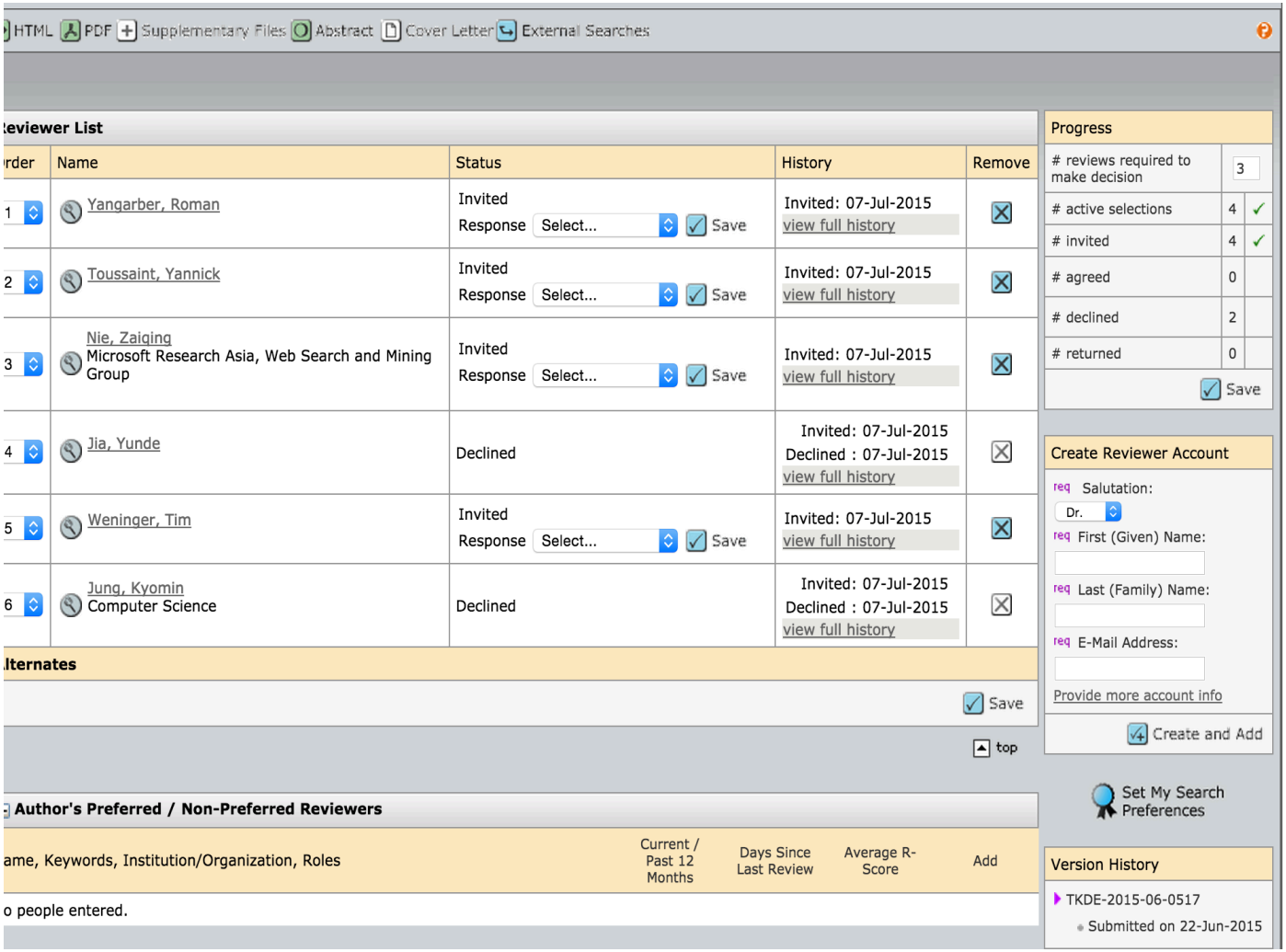
清華大學 Tsinghua University

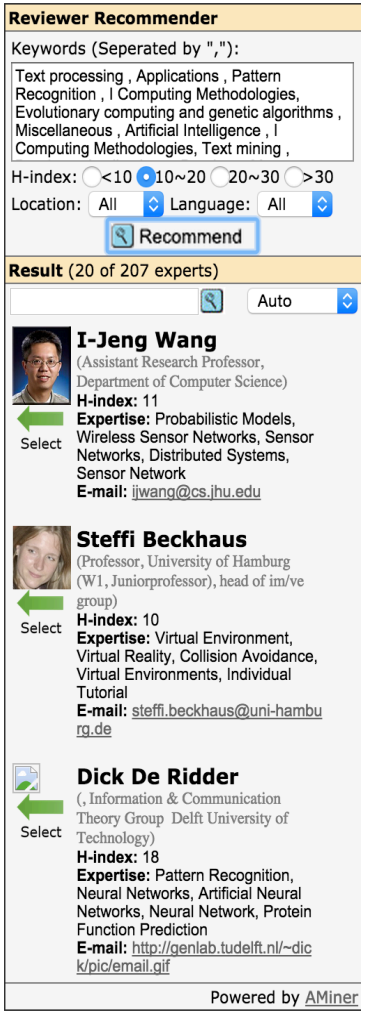
# Reviewer Suggestion




Inerest matching
COI avoiding
Load balancing
Forcast review quality

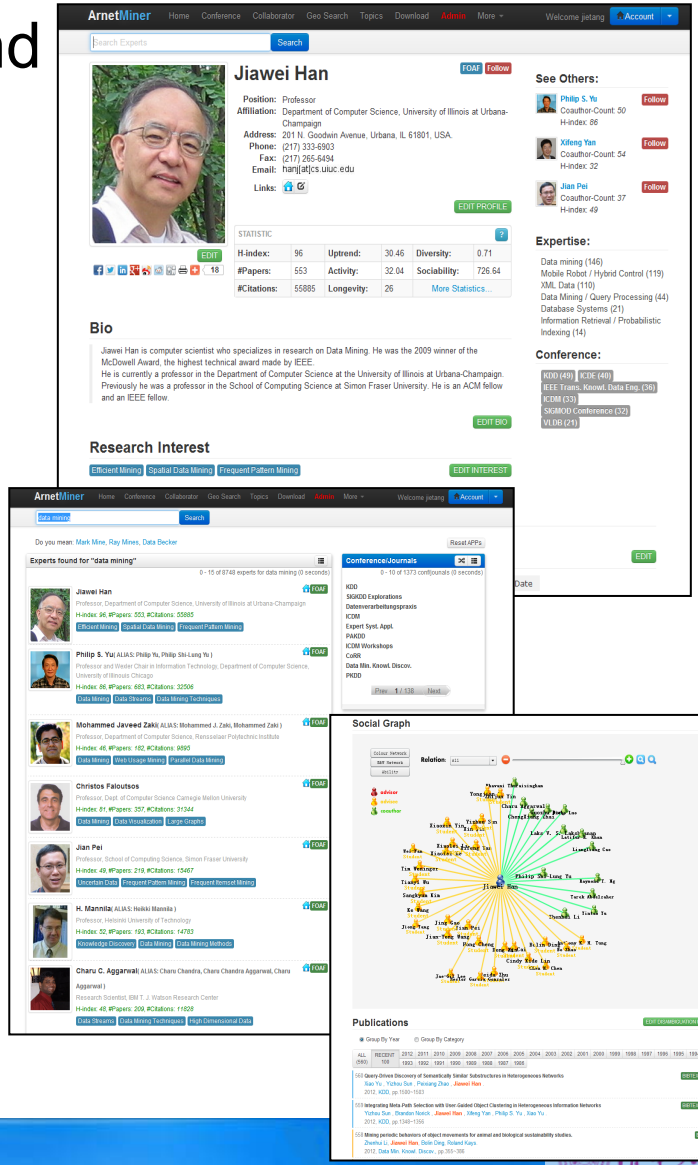
# Reviewer Suggestion



# AMiner II (ArnetMiner)

- Academic Social Network Analysis and Mining system—AMiner (http://aminer.org)
  - Online since 2006
  - **>38 million** researcher profiles
  - **>76 million** publication papers
  - **>241 million** requests
  - **>12.35 Terabyte** data
  - 100K IP access from 170 countries per month
  - 10% increase of visits per month
- Deep analysis, mining, and search



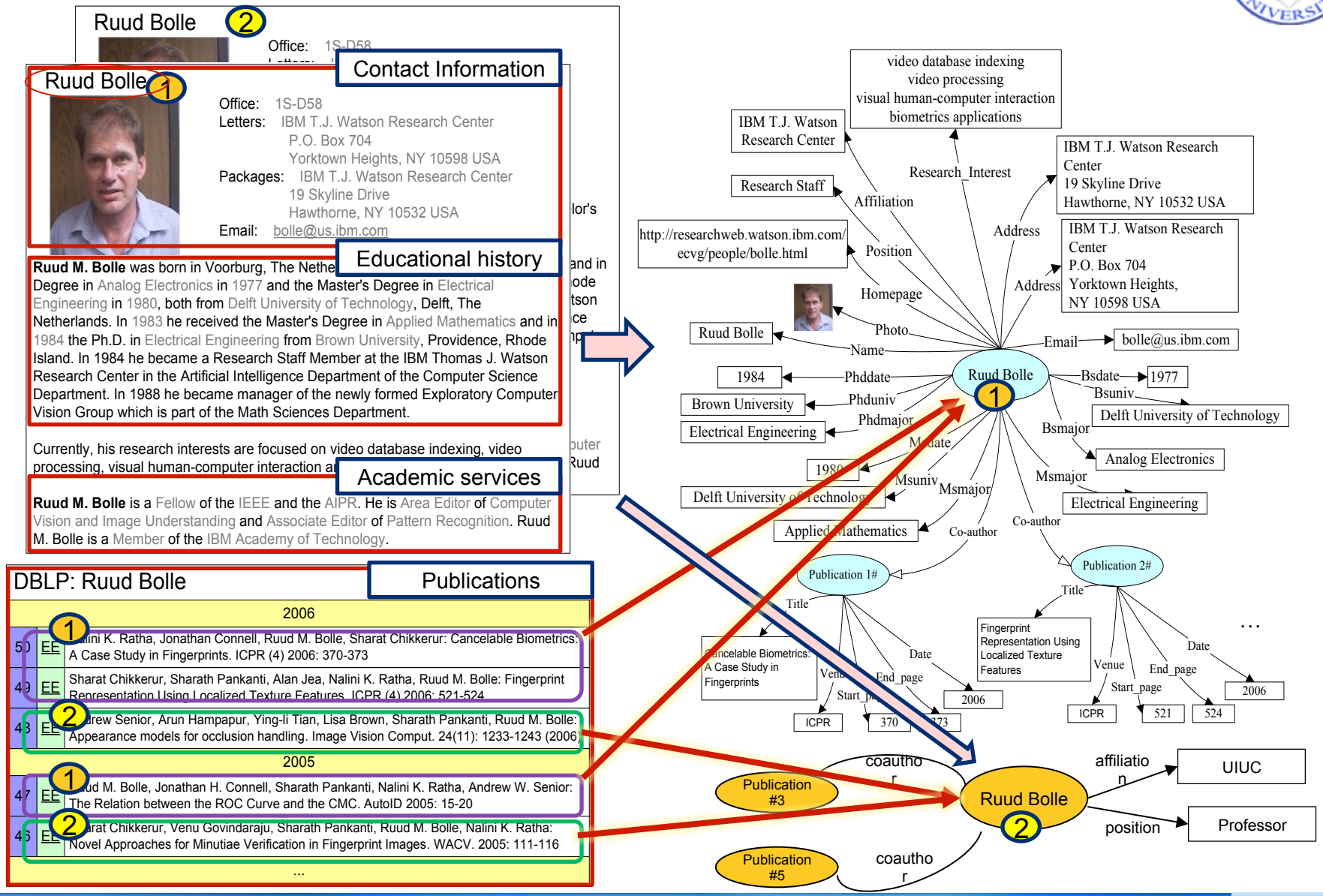
# User Distribution

**7.32 million IP from 220 countries/regions**

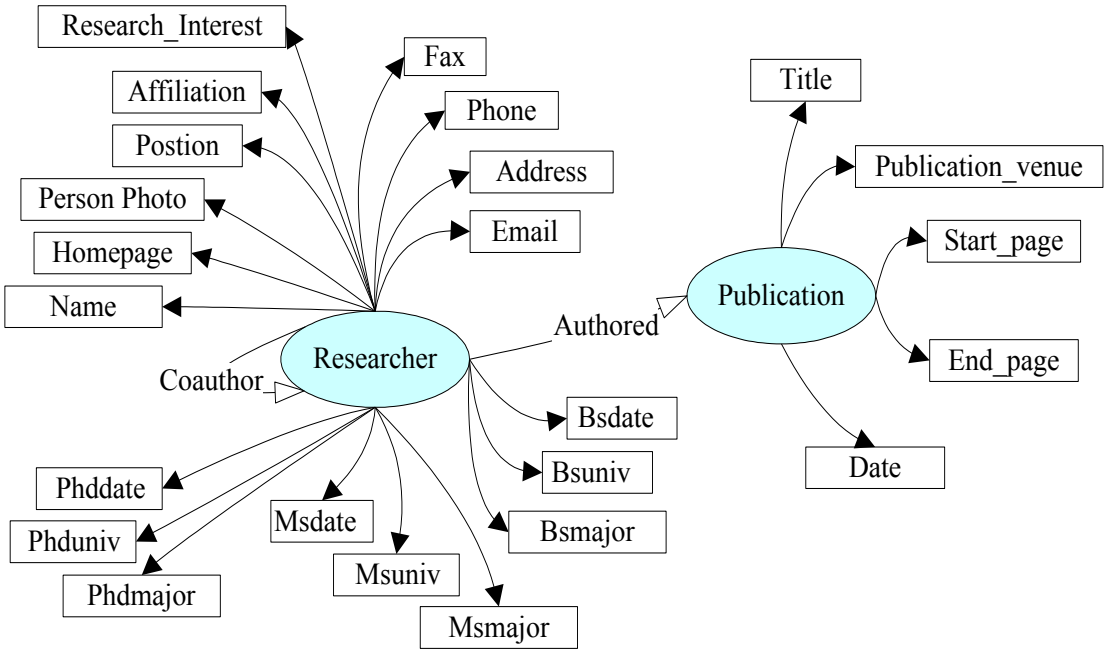
## Top 10 countries

1. USA
2. China
3. Germany
4. India
5. UK
6. Canada
7. Japan
8. Spain
9. France
10. Italy

# Motivating Example



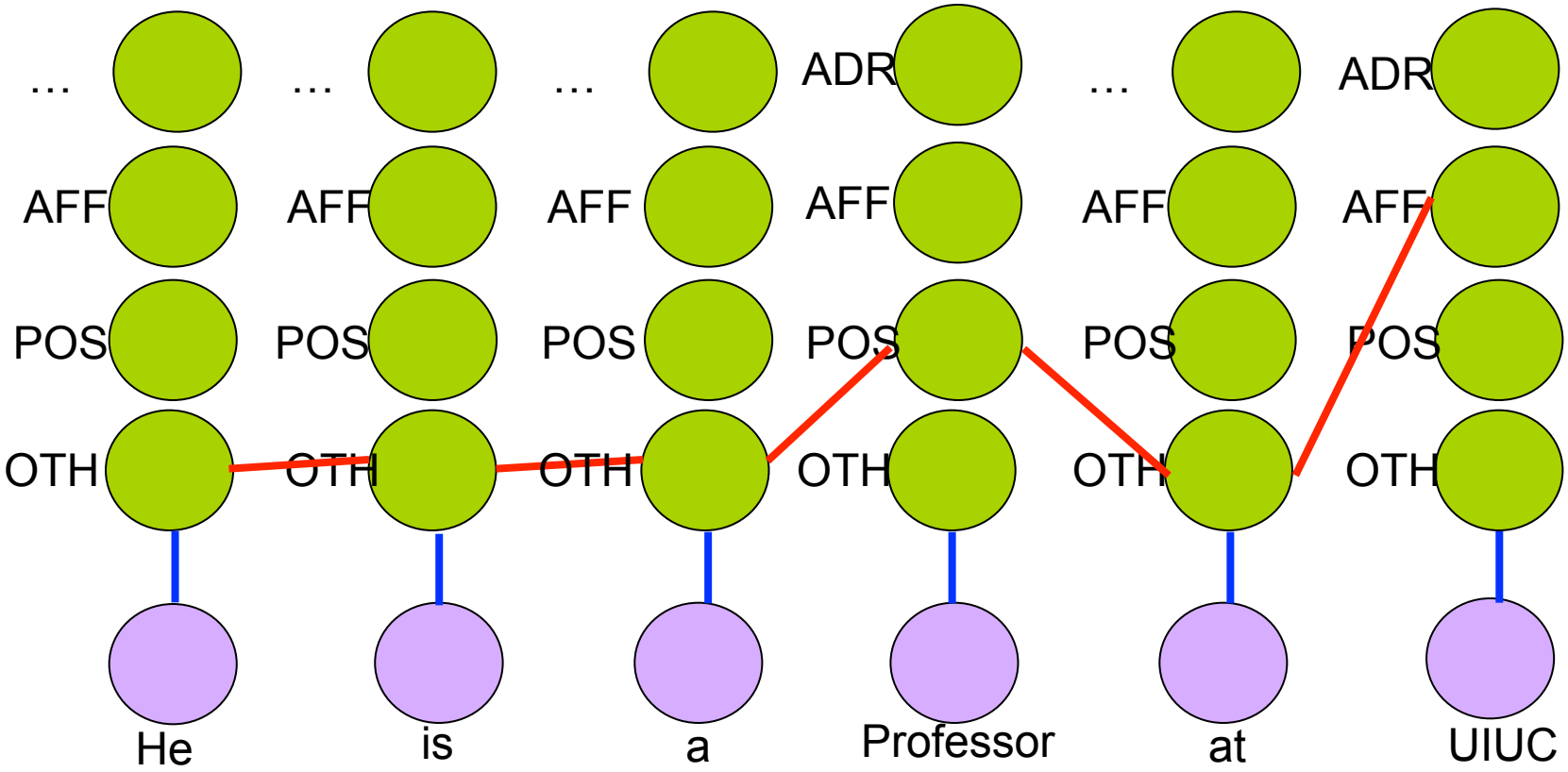
# Researcher Social Network Extraction




| 70.60% of the researchers have at least one homepage or an introducing page | |
|---|---|
| 85.6% from universities | 14.4% from companies |
| 71.9% are homepages | 28.1% are introducing pages |
| 40% are in lists and tables | 60% are natural language text |

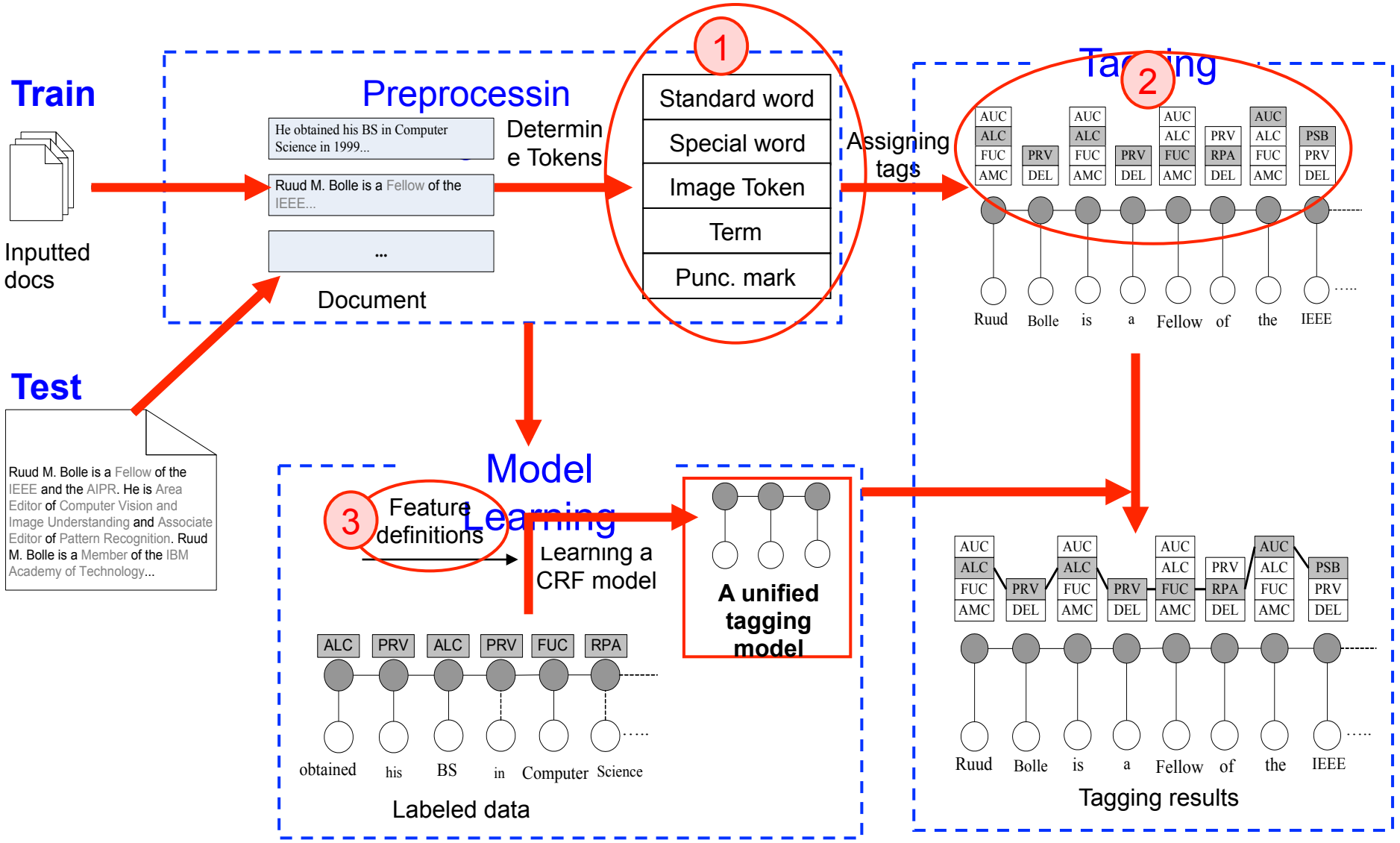
# CRFs

- Green nodes are *hidden vars,* - Purple nodes are *observations*




$$p(y \mid x) = \frac{1}{Z(x)} \exp\left( \sum_{e \in E, j} \lambda_j t_j(e, y|_e, x) + \sum_{v \in V, k} \mu_k s_k(v, y|_v, x) \right)$$

# Processing Flow for Profiling



# Profiling Results—5-fold cross validation

| Profiling Task | Unified | Unified_NT | SVM | Amilcare |
|---|---|---|---|---|
| Photo | 89.11 | 88.64 | 88.86 | 31.62 |
| Position | 69.44 | 64.70 | 64.68 | 56.48 |
| Affiliation | 83.52 | 72.16 | 73.86 | 46.65 |
| Phone | 91.10 | 78.72 | 79.71 | 83.33 |
| Fax | 90.83 | 64.28 | 64.17 | 86.88 |
| Email | 80.35 | 75.47 | 79.37 | 78.70 |
| Address | 86.34 | 75.15 | 77.04 | 66.24 |
| Bsuniv | 67.38 | 57.56 | 59.54 | 47.17 |
| Bsmajor | 64.20 | 59.18 | 60.75 | 58.67 |
| Bsdate | 53.49 | 40.59 | 28.49 | 52.34 |
| Msuniv | 57.55 | 47.49 | 49.78 | 45.00 |
| Msmajor | 63.35 | 61.92 | 62.10 | 57.14 |
| Msdate | 48.96 | 41.27 | 30.07 | 56.00 |
| Phduniv | 63.73 | 53.11 | 57.01 | 59.42 |
| Phdmajor | 67.92 | 59.30 | 59.67 | 57.93 |
| Phddate | 57.75 | 42.49 | 41.44 | 61.19 |
| **Overall** | **83.37** | **72.09** | **73.57** | **62.30** |

# Outline

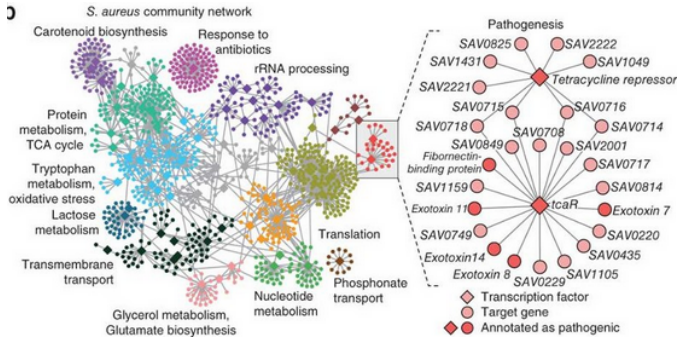
- Knowledge graph and technologies
- Big scholar knowledge base - Aminer II
- Knowledge graph building over enterprise data
- Conclusion

# Knowledge Graph over Enterprise Data

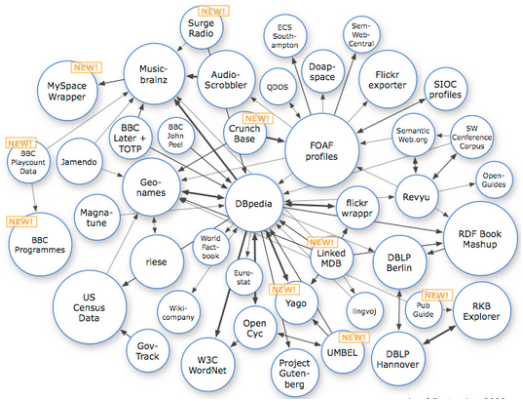
## Motivation

The current constructions of the knowledge graph are mainly from two aspects: Web, Domains, Science

Gene Ontology




LOD




There is huge demand on knowledge graph building based on internal data of enterprise

# Knowledge Graph over Enterprise Data

## Building Knowledge Graph based on Mobile Customer Care Documents

- Document Parsing based Logical Structure Extraction
- Heuristic Table Extraction
- Hierarchical Concept Extraction
- Iterative Instance Identification & Property Extraction
- Evaluation Throughout Performance

# Knowledge Graph over Enterprise Data

## Evaluations

- Document Parsing Evaluation

$$section\_precision = \frac{correct\_section}{sum\_ext\_section} \qquad section\_recall = \frac{correct\_section}{sum\_org\_section}$$

$$block\_precision = \frac{correct\_block}{sum\_ext\_block} \qquad block\_recall = \frac{correct\_block}{sum\_org\_block}$$

- Table Alignment Evaluation
  - via manual evaluation
- Knowledge Graph Evaluation
  - Coverage

# Summary

- Domain data characteristics
- Problem to be solved
- Building pipeline
- Linking with external knowledge graphs
- Visualization and Evaluation
  - On each process
  - Knowledge base evaluation
- Human interaction

# Conclusions

- There have a lot of available knowledge graphs and large scale linked data
- There have various knowledge representation and learning methods for different kinds of format sources
- knowledge graph building pipeline and toolkits within specific domains can facilitate the fast building of specific domain knowledge graph
- Making using existing knowledge graphs as possible as we can

# Thanks & Questions