Recent Trends in Digital Library Research - Mainly about JCDL

Kazunari Sugiyama (National University of Singapore)
Tadashi Nomoto (National Institute of Japanese Literature)

Information Access Symposium 2012
Special Interest Group of Information Fundamentals and Access Technologies (SIG-IFAT),
Information Processing Society of Japan (IPSJ)

Contents

- Outline of JCDL 2012 (by Sugiyama)
- Research Trends Related to Information Retrieval (by Sugiyama)
- Research Trends Related to Citation Analysis (by Prof. Nomoto)

Outline of JCDL 2012

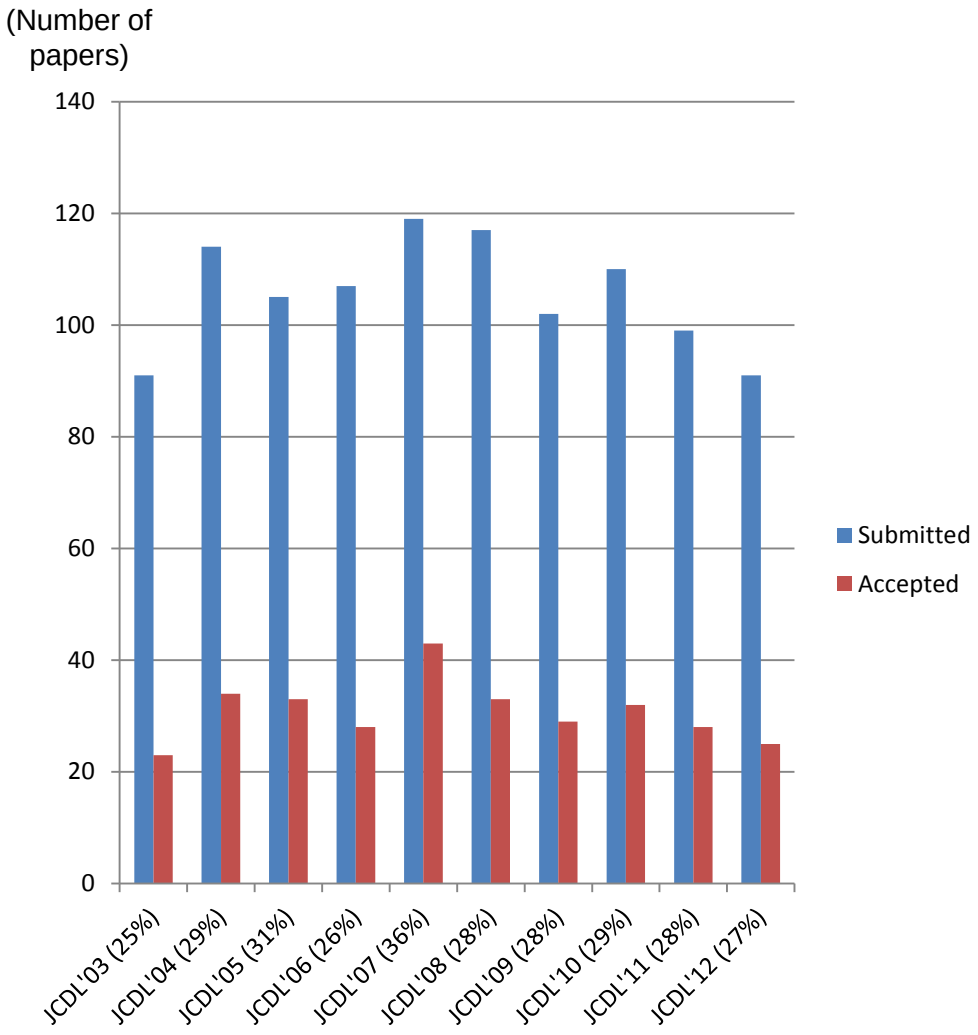
- Venue
  - George Washington University, Washington DC, US

Outline of JCDL 2012

(Cited from “The George Washington University, Foggy Bottom Campus Walking Tour”)

Outline of JCDL 2012

- Acceptance rate in JCDL‘12
  - Submitted papers: 201
  - Accepted papers:
    - Full papers 25 / 91 (27.4%)
    - Short papers 22/ 70 (31.4%)
- JCDL’13@Indianapolis, US
  - Paper due: 28th Jan, ‘13



Acceptance rate of full paper from 2003 to 2012

Doctoral Consortium in JCDL 2012

- 9 Ph.D. students
  - China (1), UK (1), US (7)
- Topics
  - Archive
  - Crowdsourcing
  - Metadata
  - User generated content
  - User search behavior
- Doctoral Consortium in JCDL ‘13
  - Paper due: 15th Apr., ‘13

Research Trends Related to IR

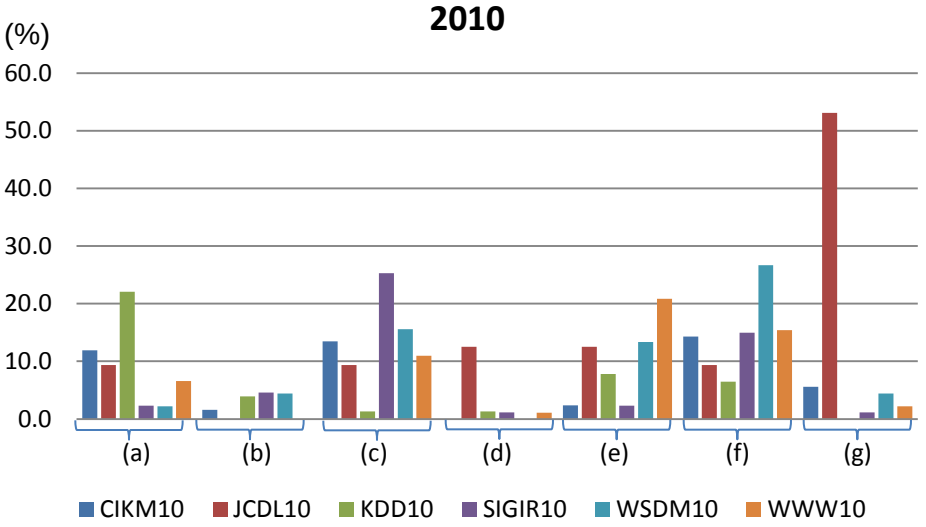
- Analyze IR related conferences based on topics that JCDL mainly addresses
  - CIKM
  - JCDL
  - KDD
  - SIGIR
  - WSDM
  - WWW

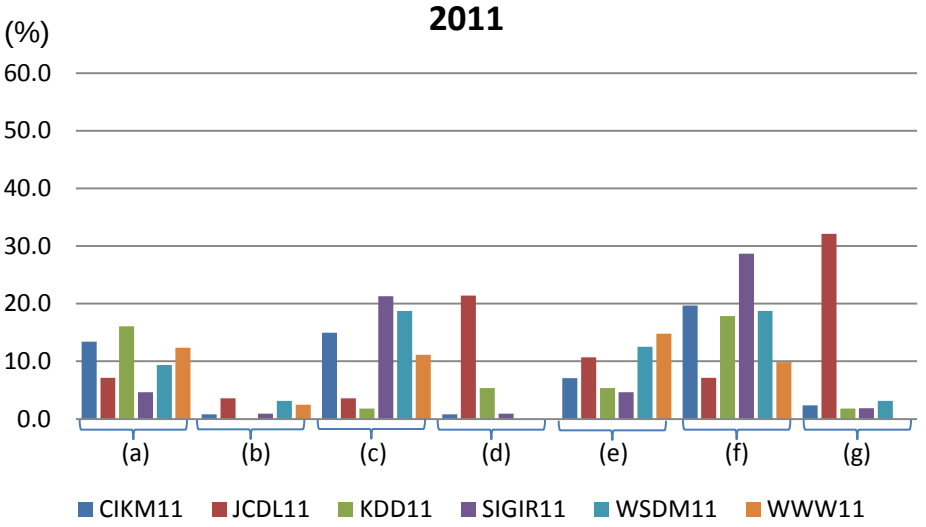
Topics in JCDL

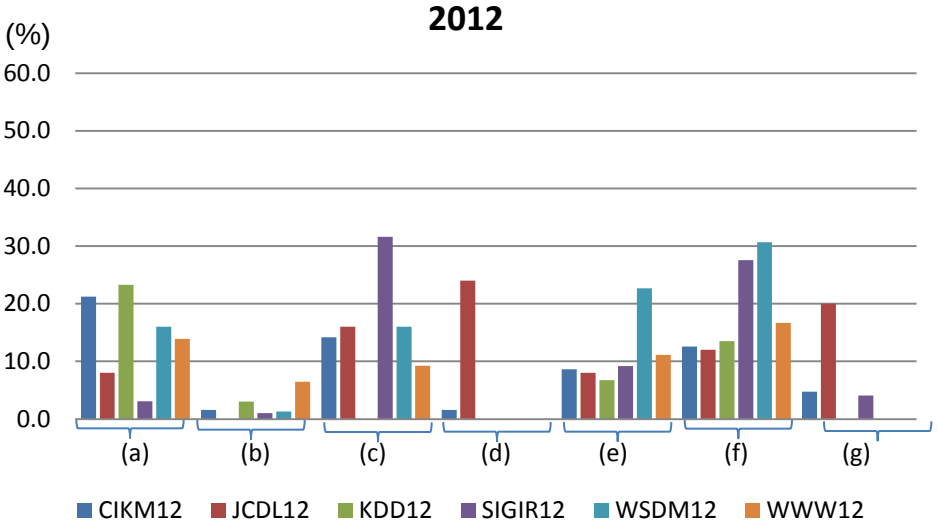
- Collaborative and participatory information environments
- Cyberinfrastructure architectures, applications, and deployments
- Data mining/extraction of structure from networked information
- Digital library and Web Science curriculum development
- Distributed information systems
- Extracting semantics, entities, and patterns from large collections
- Evaluation of online information environments
- Impact and evaluation of digital libraries and information in education
- Information and knowledge systems
- Information policy and copyright law
- Information visualization
- Interfaces to information for novices and experts
- Linked data and its applications
- Personal digital information management
- Retrieval and browsing
- Scientific data curation, citation and scholarly publication
- Social media, architecture, and applications
- Social networks, virtual organizations and networked information
- Social-technical perspectives of digital information
- Studies of human factors in networked information
- Theoretical models of information interaction and organization
- User behavior and modeling
- Visualization of large-scale information environments
- Web archiving and preservation

(Cited from http://jcdl2012.info/call-for-papers)

Research Trends from 2010 to 2012

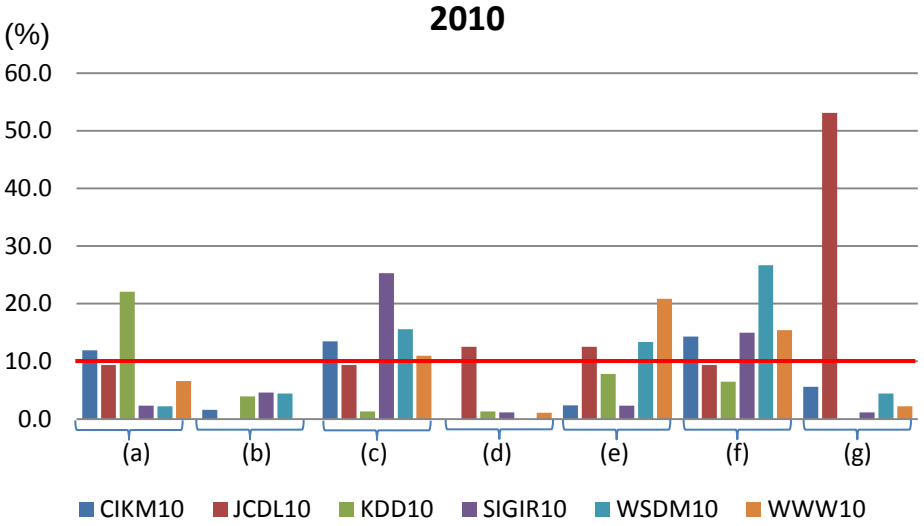


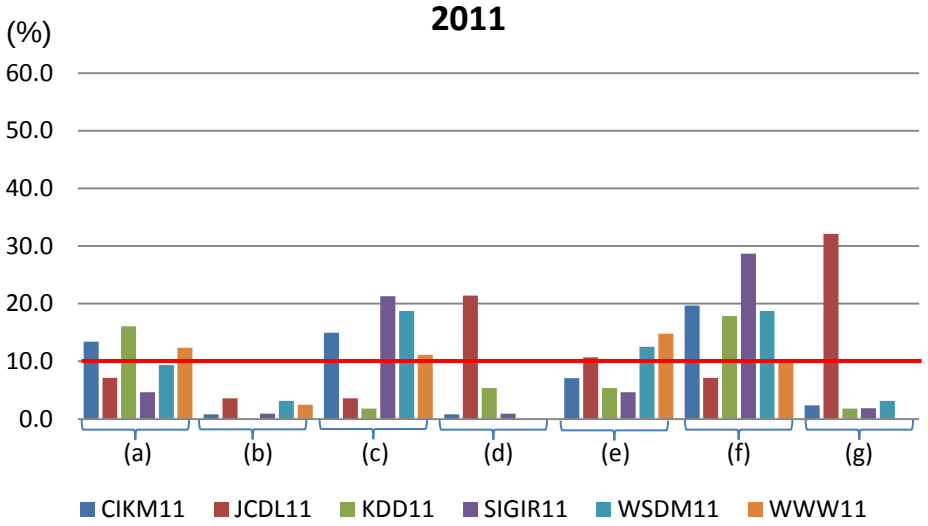


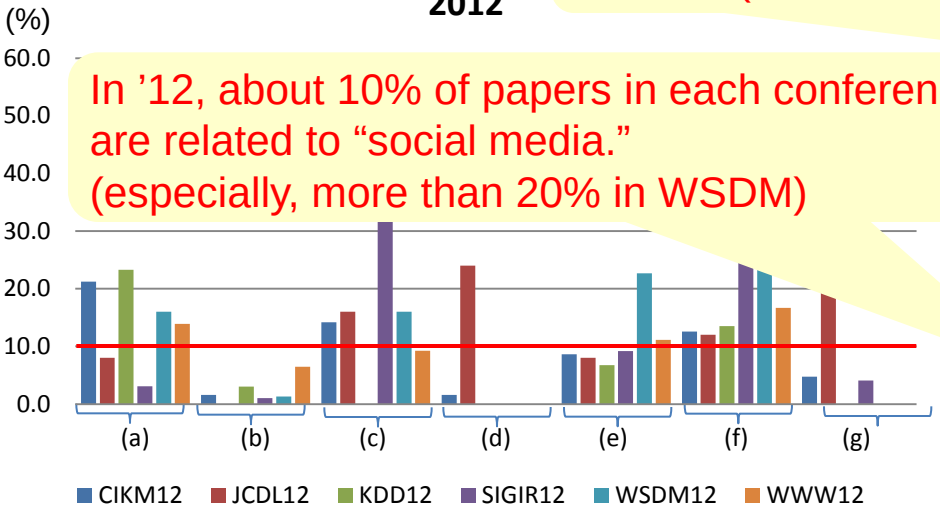


(a) Data mining/extraction of structure from networked information
(b) Linked data and its applications
(c) Retrieval and browsing
(d) Scientific data curation, citation and scholarly publication
(e) Social media, architecture, and its applications
(f) User behavior and modeling
(g) Web archiving and preservation

Research Trends from 2010 to 2012







Significantly increased in CIKM ('10, '11 -> '12), WSDM ('10 -> '11-> '12), and WWW ('10 -> '11, '12)

In '12, about 10% of papers in each conferences are related to "social media." (especially, more than 20% in WSDM)

(a) Data mining/extraction of structure from networked information

(b) Linked data and its applications

(c) Retrieval and browsing

(d) Scientific data curation, citation and scholarly publication

(e) Social media, architecture, and its applications

(f) User behavior and modeling

(g) Web archiving and preservation

“Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking” (the best paper in JCDL’12)

Hongbo Deng (University of Illinois at Urbana-Champaign)

Jiawei Han (University of Illinois at Urbana-Champaign)

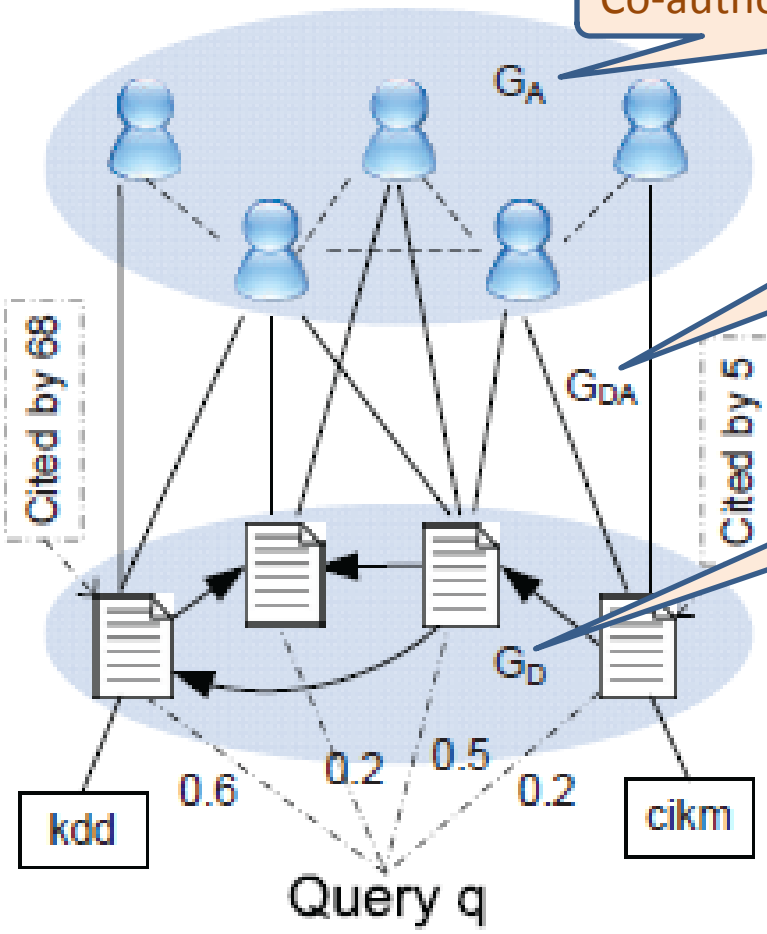
Michael R. Lyu (The Chinese University of Hong Kong)

Irwin King (The Chinese University of Hong Kong)

Outline

- Expertise ranking
  - Important in community-based question answering system and bibliography data as expertise are actively publishing useful contents
  - Model and exploit heterogeneous network together with textual content information
    - Existing approaches only take textual documents into account.
  - Formulate three types of hypotheses that capture different information in the heterogeneous network with respect to different types of edges

Heterogeneous Network in Bibliography Data



Query: "network analysis"

Existing approaches

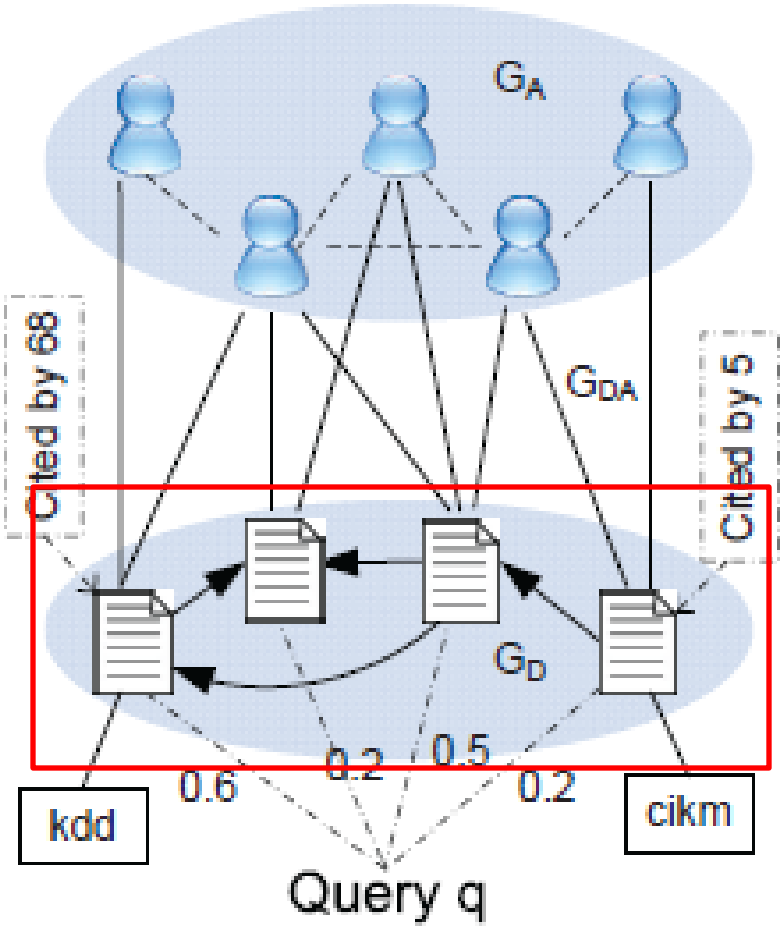
P(Deng | "network analysis") = 0.73
[G_D]

Proposed approaches

P(Deng | "network analysis") = 0.81
[G_D, G_{DA}, G_A]

(Cited from Figure 1 in "H. Deng et al. "Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking")

Hypothesis to Incorporate Different Types of Graphs

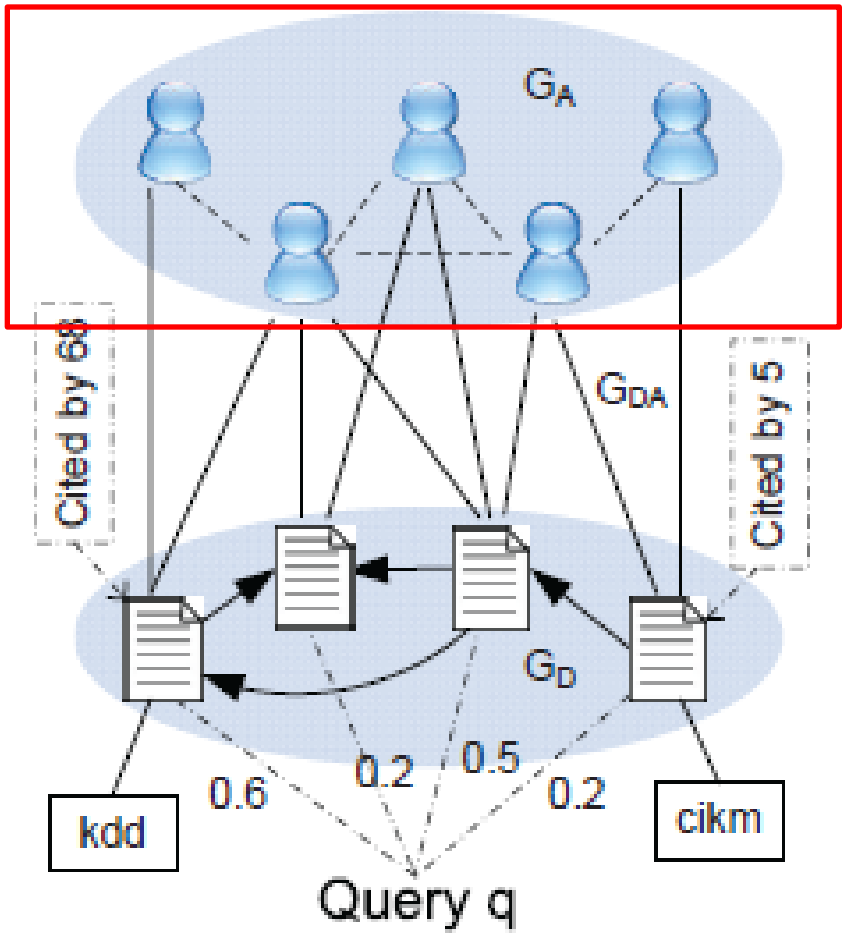


- Document Consistency
  - The neighbors of a target document are considered similar to it.

Make the difference in relevance scores between neighbor documents small for a given query

(Cited from Figure 1 in "H. Deng et al. "Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking")

Hypothesis to Incorporate Different Types of Graphs

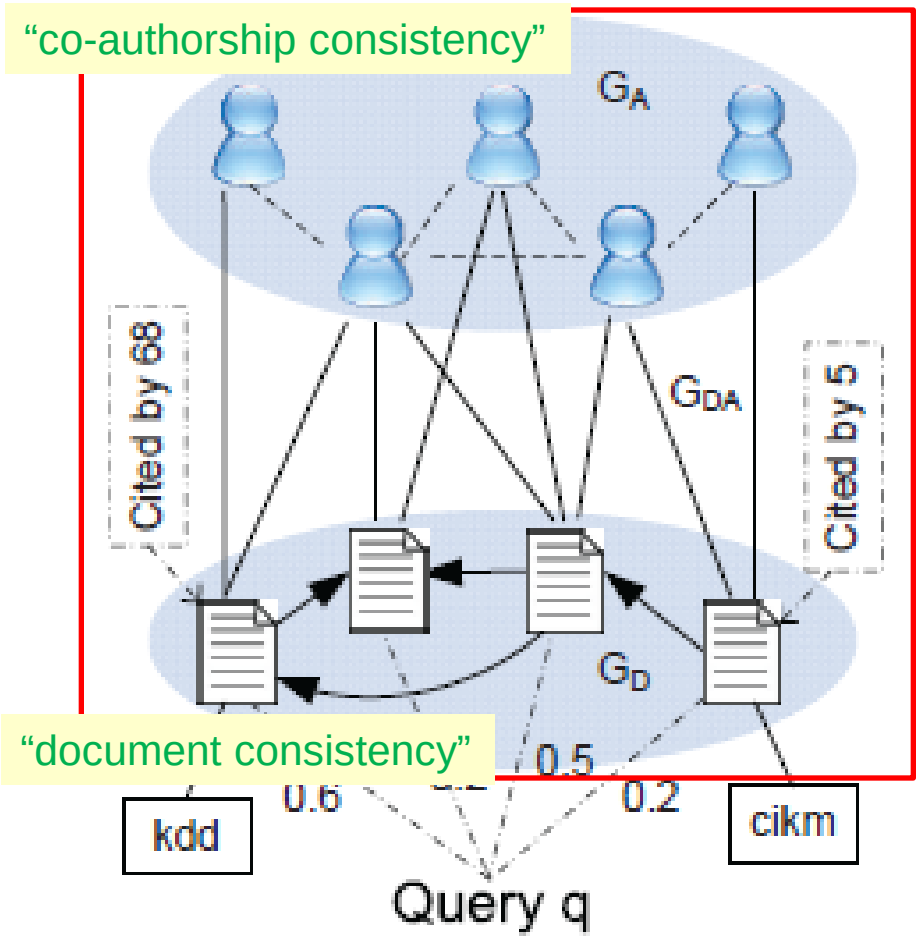


- Co-authorship Consistency
  - If two persons have many co-authored papers, then their expertise should be similar.

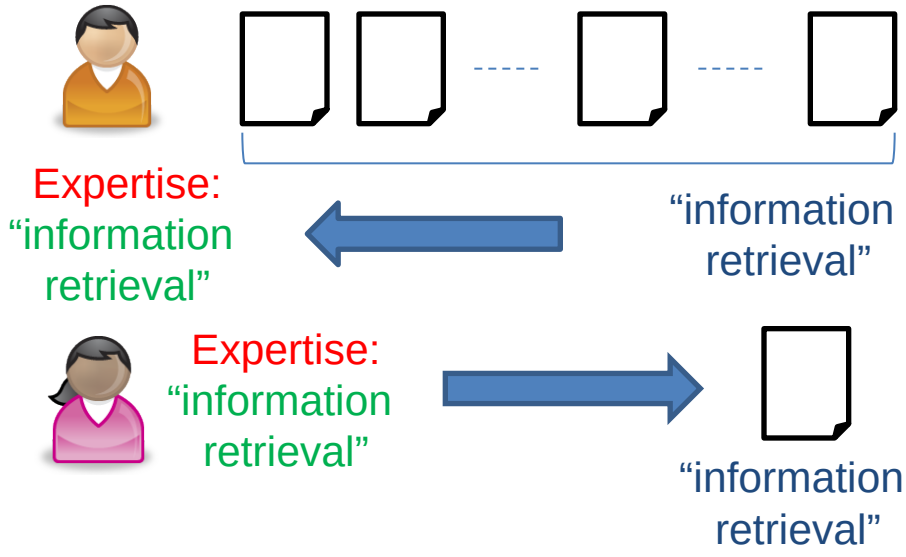
Make the expertise scores of candidates to be closer if they co-authored more papers for a given query

(Cited from Figure 1 in "H. Deng et al. "Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking")

Hypothesis to Incorporate Different Types of Graphs



- Document-Author Consistency
  - The expertise of a researcher is consistent with that of documents he/she published.



Integrate "document consistency and "co-authorship consistency" into expertise ranking for a given query.

(Cited from Figure 1 in "H. Deng et al. "Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking")

Experiments

- Dataset
  - DBLP bibliography data
    - 1,152,512 papers, 695,906 authors
  - Manually created 20 query topics

    "Information Extraction," "Intelligent Agents," "Machine Learning," "Natural Language Processing," "Planning," "Semantic Web," "Support Vector Machine," "Boosting," "Ontology Alignment," "Probabilistic Relevance Model," "Information Retrieval," "Language Model for Information Retrieval," "Face Recognition," "Semi Supervised Learning," "Reinforcement Learning," "Kernel Methods," "Privacy Preservation," "Skyline," "Sensor RFID data management," "Stream"
  - Maximum number of experts
    - "Boosting" (56)
  - Minimum number of experts:
    - "Language Model for Information Retrieval" (12)
  - Average number of experts per query
    - 26.9

Experiments

- Results

	P@5	P@10	P@20	MRR
Baseline	0.700	0.670	0.498	0.900
Proposed method	**0.840**	**0.720**	**0.570**	**1.000**

Baseline: Estimates the expertise of candidate based on both the relevance and quality of associated documents

Experiments

- Example of expert search

The top 10 experts for query “Probabilistic Relevance Model”

Baseline	Proposed method
Norbert Fuhr	**Norbert Fuhr**
Stephen E. Robertson	**Stephen E. Robertson**
Friedrich Gebhardt	**W. Bruce Croft**
M. E. Maron	**M. E. Maron**
J. L. Kuhns	**ChengXiang Zhai**
ChengXiang Zhai	**C. J. van Rijsbergen**
C. J. van Rijsbergen	Friedrich Gebhardt
Azadeh Shakery	**J. L. Kuhns**
W. Bruce Croft	**Chris Buckley**
Victor Lavrenko	**William S. Cooper**

(* Bold font: Relevant experts)