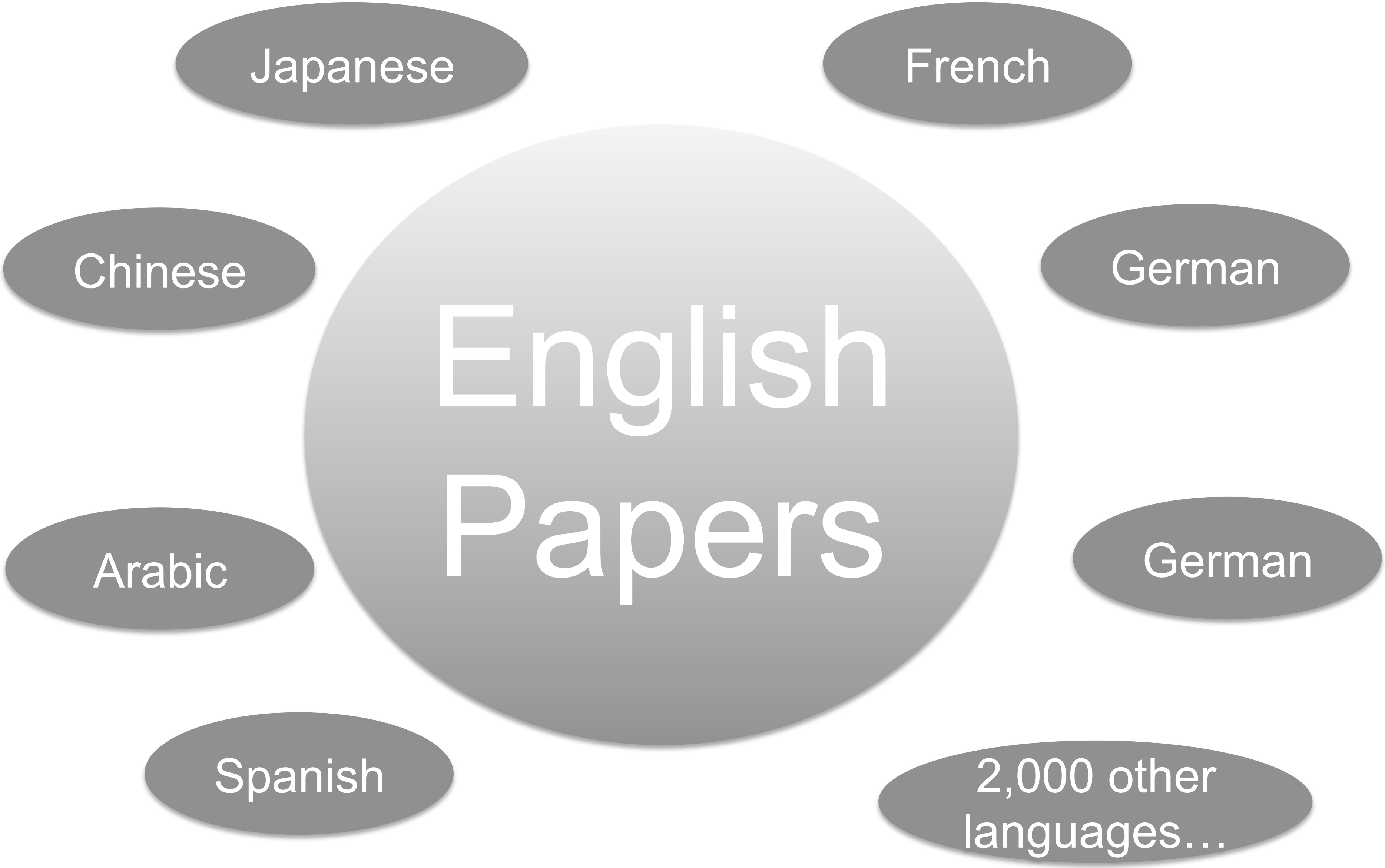
Interlinking Cross Language Metadata Using Heterogeneous Graphs and Wikipedia

XIAOZHONG LIU Indiana University Bloomington
MIAO CHEN Indiana University Bloomington
JIAN QIN Syracuse University

DC2014, Austin, TX, USA, October 9, 2014

Recent research has revealed that most articles published in top US accounting journals come from institutions based in the US or a small number of other English-speaking countries (Jones and Roberts, 2005)… most recognized academic journals are located in the US or other English-speaking countries, with the consequence that they only accept papers in English. Even for journals with a more international basis, English is the only permitted language… (Raffournier & Schatt, 2010).

English Papers



English Papers

Japanese

French

Chinese

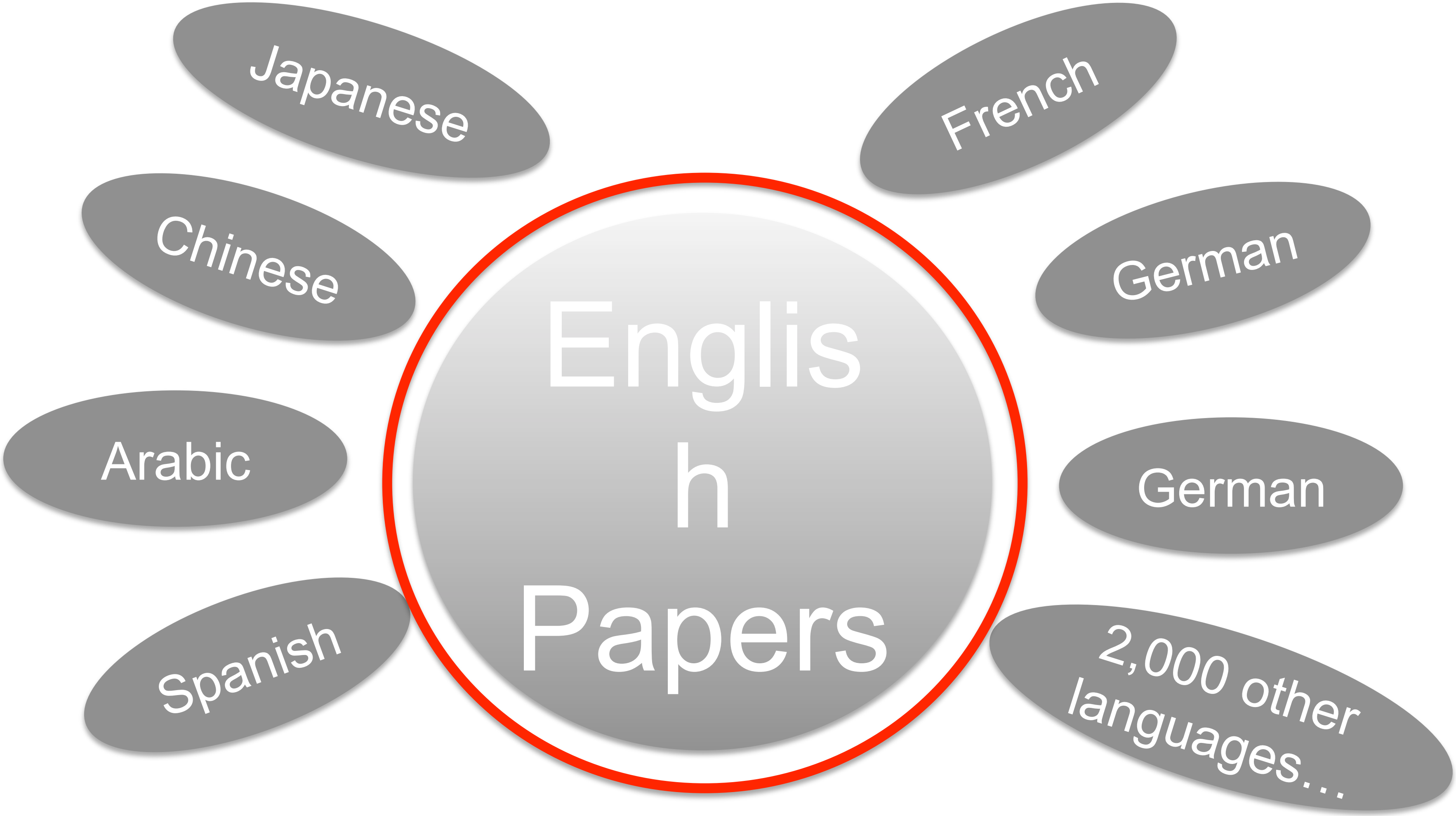
German

Arabic

German

Spanish

2,000 other languages…



Japanese

French

Chinese

German

Arabic

English Papers

German

Spanish

2,000 other languages…

How to break this language barrier?
Users from different countries can easily access those English papers…

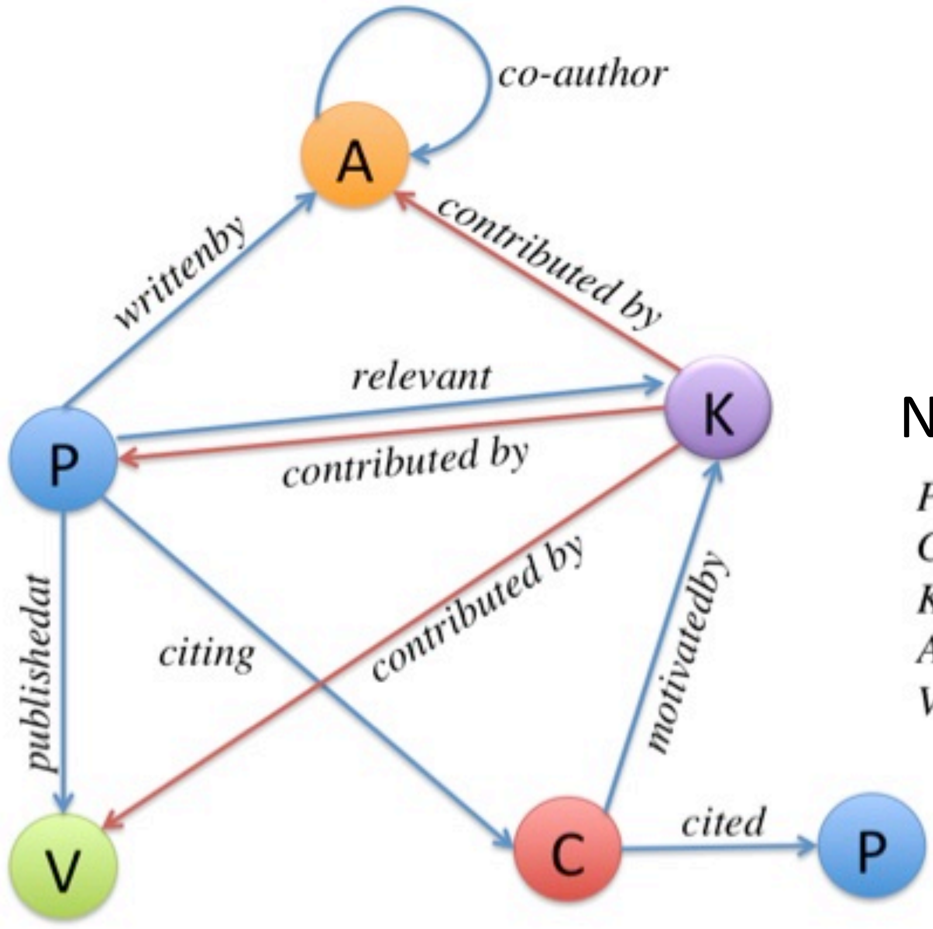
Recent research has revealed that most articles published in top US accounting journals come from institutions based in the US or a small number of other English-speaking countries (Jones and Roberts, 2005)… most recognized academic journals are located in the US or other English-speaking countries, with the consequence that they only accept papers in English. Even for journals with a more international basis, English is the only permitted language… (Raffournier & Schatt, 2010).

Limitations

- Lack of *citation relationships,* e.g., paper A (in Language A) cites English papers

- Difficulty in *personalizing* user profiles, i.e., how to construct a user profile (in Language A)

All the papers written in a specific language contain nodes and edges



Node type

P: Paper
C: Citation
K: Keyword (topic)
A: Author
V: Venue

Edges:

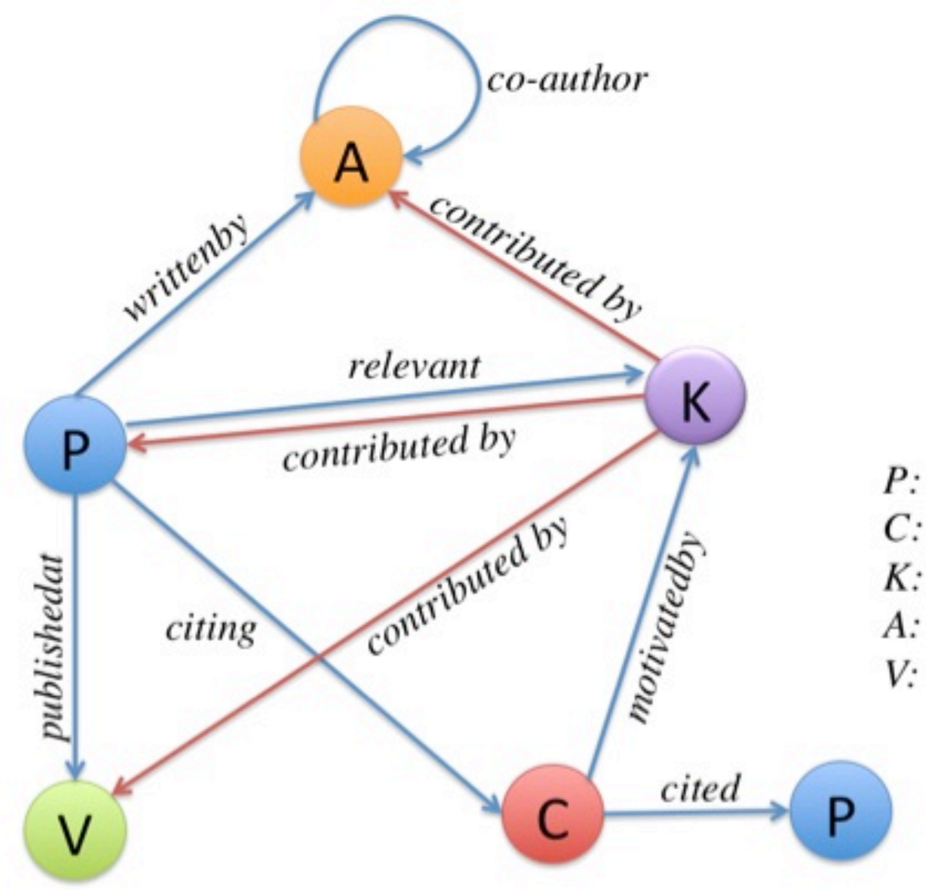
- written by
- contributed by
- published by
- relevant
- motivated by
- citing

Edges have types

1. *P→A* : a paper is written by an author;
2. *P→V* : a paper is published in a venue;
3. *P→K* : a paper or publication is relevant to a keyword;
4. *P→P* : a publication cites or links to publications;
5. *K→ P* : a keyword (topic) is assigned to publications;
6. *K→A* : a keyword (topic) is assigned by authors; and
7. *K→V :* a keyword (topic) is assigned to venues

Random Walks based on meta-paths

Scenario 1: Recommend Paper (P) to Author (A) based on possible paths



I: A* --> A --> $P^{?}$

II: A* --> P --> C --> $P^{?}$

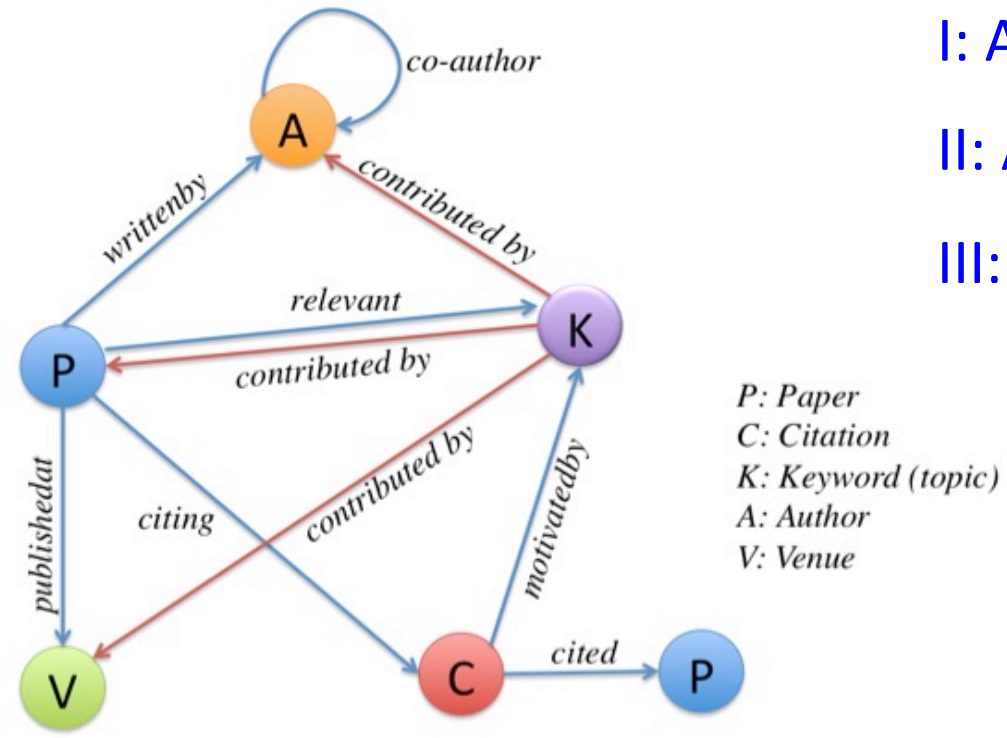
III: A* --> A --> P --> C --> $P^{?}$

IV: A* --> P --> C --> P --> V --> $P^{?}$

P: Paper
C: Citation
K: Keyword (topic)
A: Author
V: Venue

Random Walks based on meta-paths

Scenario 2: Suggest Keyword (K) to Author (A) based on possible paths



I: A* --> A --> K$^{?}$

II: A* --> P --> C --> P --> K$^{?}$

III: A* --> V --> P --> C --> P --> K$^{?}$

Different meta-paths are integrated into a unique recommendation model.

Proven in citation recommendation for one DL in a single language

Can meta-paths be used to bridge metadata in different languages?

(Liu, Guo, Yu, and Sun, 2014; Liu, Yu, Guo, Sun, and Gao, 2014)

Research questions

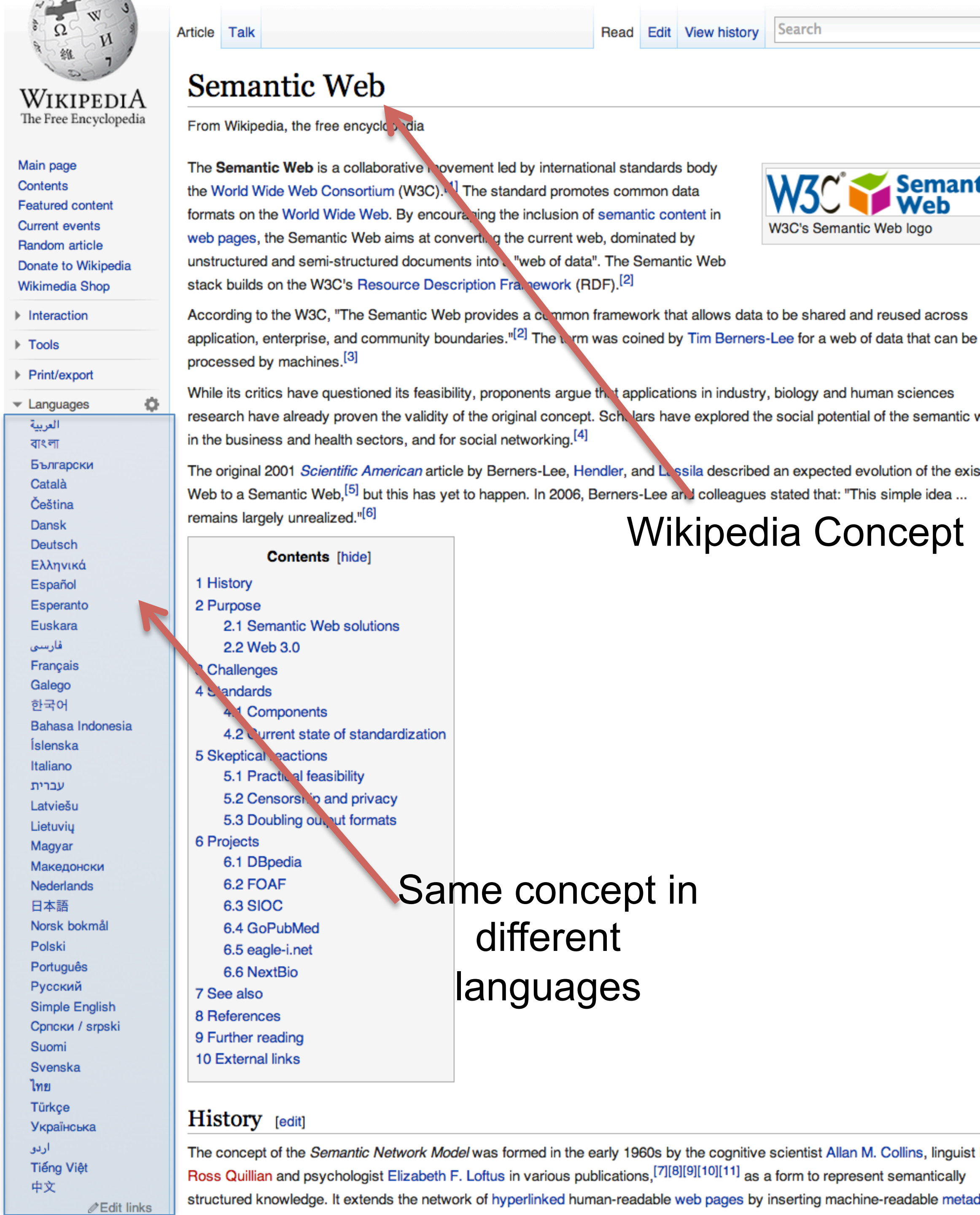
Given that two digital libraries, DL1 (in language 1) and DL2 (in language 2), have no direct connections by way of citations or authors:

- **RQ1: How can metadata for publications in DL1 and DL2 be bridged through language equivalents such as topics (keywords), authors, and venues?**

- **RQ2: How can recommendations be made for resources from DL1 to DL2?**

Wikipedia provides a source to link concepts across different languages.

All the Wikipedia concepts are **interconnected via hyperlinks and categories.**



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction

Tools

Print/export

Languages

العربية
বাংলা
Български
Català
Čeština
Dansk
Deutsch
Ελληνικά
Español
Esperanto
Euskara
فارسی
Français
Galego
한국어
Bahasa Indonesia
Íslenska
Italiano
עברית
Latviešu
Lietuvių
Magyar
Македонски
Nederlands
日本語
Norsk bokmål
Polski
Português
Русский
Simple English
Српски / srpski
Suomi
Svenska
ไทย
Türkçe
Українська
اردو
Tiếng Việt
中文

Edit links

Article Talk Read Edit View history Search

Semantic Web

From Wikipedia, the free encyclopedia

The **Semantic Web** is a collaborative movement led by international standards body the World Wide Web Consortium (W3C).[1] The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data". The Semantic Web stack builds on the W3C's Resource Description Framework (RDF).[2]

W3C Semantic Web

W3C's Semantic Web logo

According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries."[2] The term was coined by Tim Berners-Lee for a web of data that can be processed by machines.[3]

While its critics have questioned its feasibility, proponents argue that applications in industry, biology and human sciences research have already proven the validity of the original concept. Scholars have explored the social potential of the semantic w in the business and health sectors, and for social networking.[4]

The original 2001 *Scientific American* article by Berners-Lee, Hendler, and Lassila described an expected evolution of the exis Web to a Semantic Web,[5] but this has yet to happen. In 2006, Berners-Lee and colleagues stated that: "This simple idea ... remains largely unrealized."[6]

Wikipedia Concept

Contents [hide]

Same concept in different languages

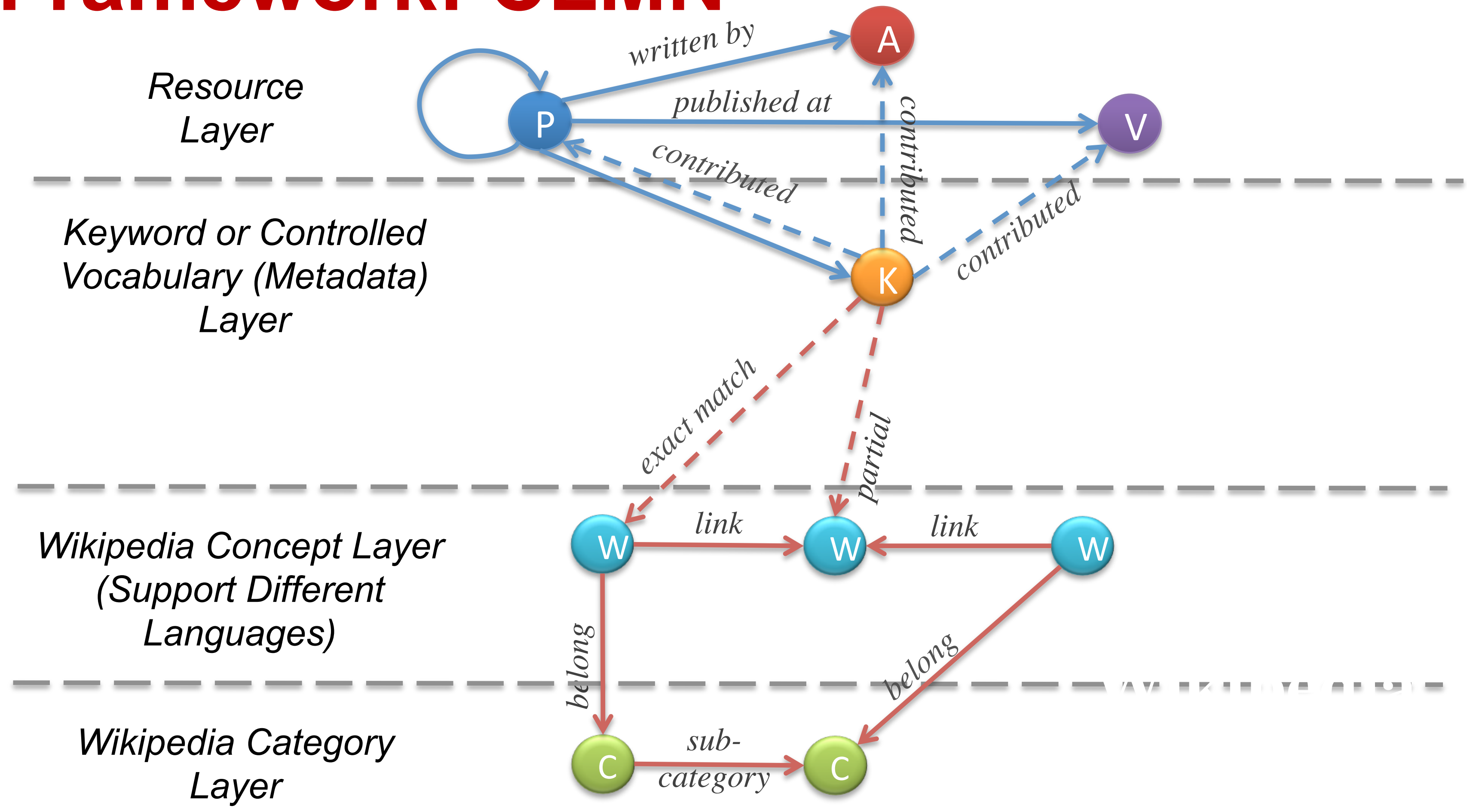
History [edit]

The concept of the *Semantic Network Model* was formed in the early 1960s by the cognitive scientist Allan M. Collins, linguist Ross Quillian and psychologist Elizabeth F. Loftus in various publications,[7][8][9][10][11] as a form to represent semantically structured knowledge. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metad

A solution for the problem:

Cross-Language
Metadata Network
(CLMN)

Framework: CLMN



Resource Layer

Keyword or Controlled Vocabulary (Metadata) Layer

Wikipedia Concept Layer (Support Different Languages)

Wikipedia Category Layer

First step: a Single-Language Metadata Network (SLMN) is built for a monolingual digital library or repository.

Second step: the SLMN will be mapped to Wikipedia concepts and subject categories to create Cross-Language Metadata Networks (CLMN).

Methods

DL1 = ACM Digital Library (in English)
DL2 = WanFang Digital Library (in Chinese)

Interim: Wikipedia 2014 May Dump

Preliminary experiment:

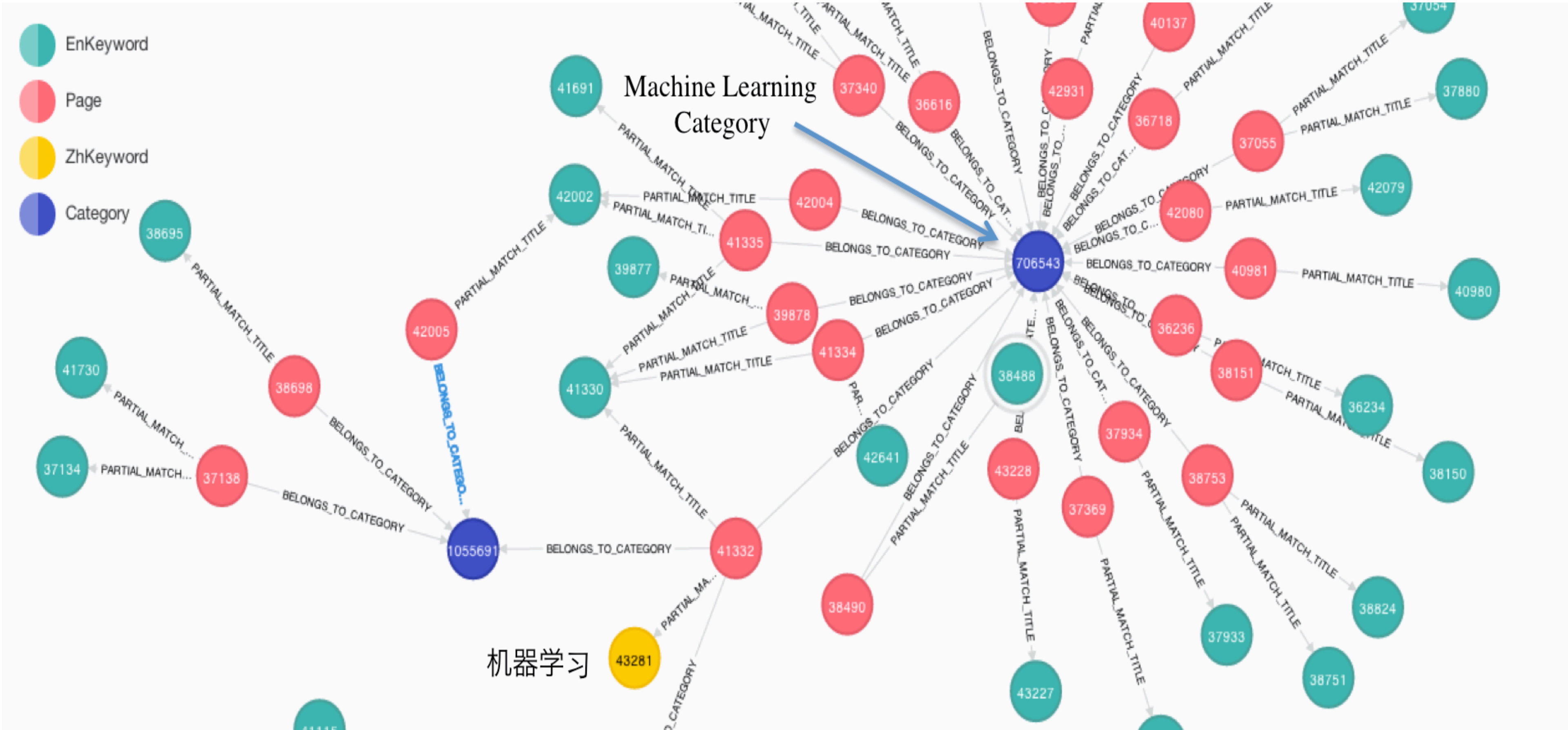
Input: a Chinese query topic
Output: related English topics

Two random walk functions:

1. [Chinese Keyword] → [Wikipedia Concept] ←[English Keyword]
2. [Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword]

Experiment results

Query: 机器学习 (Machine learning)
ACM topics related to this topic via Wikipedia page and Wikipedia categories



Machine Learning Category

机器学习

[Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword] (26 results)

①CK:机器学习 → WP:machine_learning → WC:Machine_learning ← WP:cluster_analysis ←EK:cluster_analysis

②CK:机器学习→ WP:machine_learning → WC:Machine_learning ← WP:expectation_maximization_algorithm ← EK:em_algorithm

③CK:机器学习→ WP:machine_learning → WC:Cybernetics ← WP:complex_systems ← EK:complex_systems

④CK:机器学习→ WP:machine_learning → WC:Machine_learning ← WP:reinforcement_learning ← EK:reinforcement_learning

⑤CK:机器学习→ WP:machine_learning→WC:Machine_learning ←WP:pattern_recognition ←EK:pattern_recognition

⑥CK:机器学习→WP:machine_learning→WC:Machine_learning ←WP:formal_concept_analysis ←EK:concept_analysis

Potential applications

- Automatically generate cross-language vocabularies and convert them to Linked Data format
- Recommend resources across repositories and languages based on:
  - author ID (on a SLMN)
  - keyword (on a SLMN)
  - venues (venue recommendation) or
  - expert (author recommendation)

Potential applications with the

Cross-Language Metadata Network

Language 1

Recommend

Language 2

Author	Publications
Author	Related authors
Author	Venues
Topic	Related topics
Venue	Potential reviewers

Future research

- A novel approach to generate cross-language metadata and connections
- Larger-scale experiment with evaluation by computer programs and human users
  - Validity
  - Reliability
  - Usefulness

References cited

- Liu, X., Guo, C., Yu, Y., and Sun, Y. (2014) Meta-Path-Based Ranking with Pseudo Relevance Feedback on Heterogeneous Graph for Citation Recommendation, Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM).

- Liu, X., Yu, Y., Guo, C., Sun, Y., and Gao, L. (2014) Full-Text based Context-Rich Heterogeneous Network Mining Approach for Citation Recommendation, Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL).