1st International Workshop on Standards and Technologies in Multimedia Archives and Records (STAR)

Million Book Digital Library Project - From Book to Multimedia

Huang Tiejun, and

Wen Gao, Yonghong Tian, Yizhou Wang ,Yuanning Li, etc.

Institute for Digital Media, Peking University, China

26-27th April 2010, EPFL, Lausanne, Switzerland

Content

- **Million Book Project**

- **Book Retrieval**
  - **Multilingual Access**
  - **Knowledge Organization and Retrieval**
  - **Illustration Retrieval**

- **Multimedia**
  - **Text-Image Retrieval**
  - **Video Annotating**
  - **Navigating Chinese Scrolls Fine Art**

Motivation: A Universal Digital Library

- Access to all human knowledge anytime, anywhere

- Including access, query, and print

- Any book, magazine, newspaper, painting, image, music, video, or reference document

Co-Principal Investigators

- ## Dr. Raj Reddy
  - Herbert A. Simon University Professor of Computer Science and Robotics of Carnegie Mellon University
- ## Dr. Ching-chih Chen
  - Professor and Former Associate Dean at Graduate School of Library and Information at Simmons College
- ## Prof. Qin-Ping Zhao
  - Vice President of Ministry of Education of China, Professor of Computer Science of Beijing University of Aeronautics & Astronautics
- ## Dr. Wen Gao
  - Deputy President and Professor of Computer Science, Graduate School of Chinese Academy of Sciences, as Principal Investigators

To be Millions

- 1 Million between China and US

- 1 Million between India and US

- And more: Egypt…

First Meeting
Aug.20-27,2001,CMU

Second Meeting
Mar.22-31,2002,China

The 2nd Meeting for Million Book Digital Library

Third Meeting
Jan.02-10,2003,India

International Conference on Universal Digital Library (ICUDL)

- 2005 China. Zhejiang University Oct.2006
- 2006 Egypt. Bibliotheca Alexandrina. Nov17-19
- 2007 USA Carnegie Mellon University. Nov2-4
- 2008 India IIIT-A. Oct31~Nov2
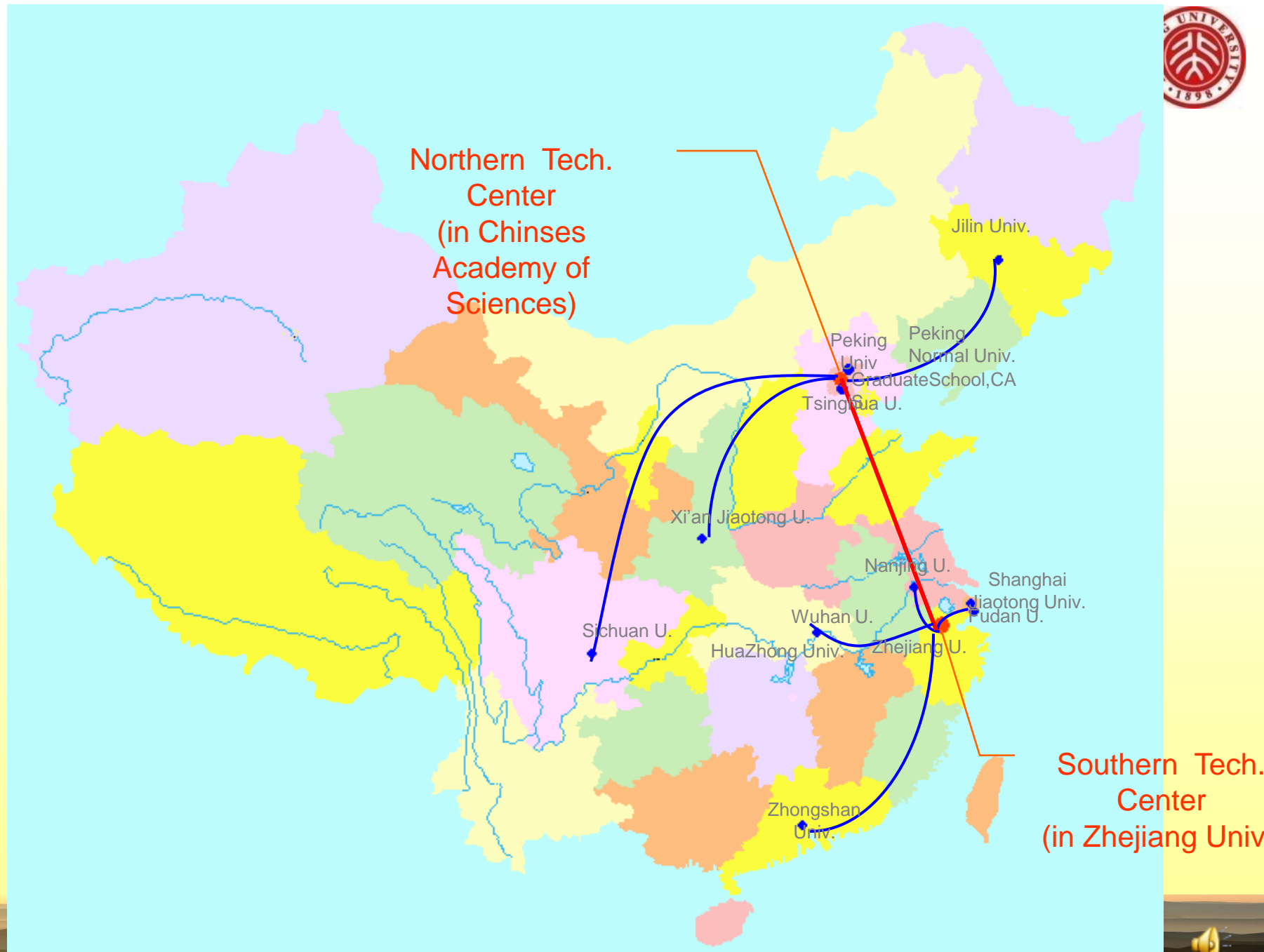- 2009 USA Pittsburgh. Nov6~8

China Government Support

- The "211" Office of MOE responsible for the program's organization and supervision

- SDPC: State Development and Planning Commission
  - 70 million RMB financial support to the project
  - 10 million RMB from ten Chinese universities involvoed
  - Fund was there in mid of 2003
  - Publicly announced by MoE in Sep.2004

- Research support
  - MOST: Ministry of Science and Technology
  - CAS: Chinese Academy of Science

Northern Tech. Center
(in Chinses Academy of Sciences)

Southern Tech. Center
(in Zhejiang Univ

Jilin Univ.

Peking Univ
Peking Normal Univ.
GraduateSchool,CA
Tsinghua U.

Xi'an Jiaotong U.

Nanjing U.

Shanghai Jiaotong Univ.
Fudan U.

Wuhan U.

Sichuan U.

HuaZhong Univ.
Zhejiang U.

Zhongshan Univ.

©北京大学黄铁军2010

Chinese collections

- Ancient Book(before 1911) (about 1/5)
  - About 100 thousand
  - Convener: Nanjing U and Peking U.
- Books during Republic of China (1911-1949)
  - About 100 thousand
  - Convener: Fudan U.
- Thesis and dissertations
  - From the Top 14 China Univ.
  - Waiting an official agreement on copyright from MoE
- Books during P R China (1949-)
  - Convener: Zhejiang U.
- More than 850 thousand finished
  - Responsible by South Tech. Center @ Zhejiang University

Mission of Northern Tech. Center

- Technical support for English book digitization
- English book management and archive
  - Exchange with southern center

- As the International resource exchange center for Million Book Project (China Site for Million book project)

- To develop a multilingual, multimedia and multimodal DL platform to provide advanced knowledge service

Digital Book profile

- Image
  - High definition (600dpi)
  - Long term archive, Book Reprint
- Metadata: exact, enough
  - MARC/CNMARC
  - Dublin Core
  - Structural metadata：METS
- Two versions: for archive, for release

Storage

- Scanning images: 80TB
  - 1 million books * 375 pages/book * 100KB/page * 2 images/page = 80TB。
- Metadata and text: about 1TB
- Multimedia content: 10TB
  - fine art, audio, video, culture bequest
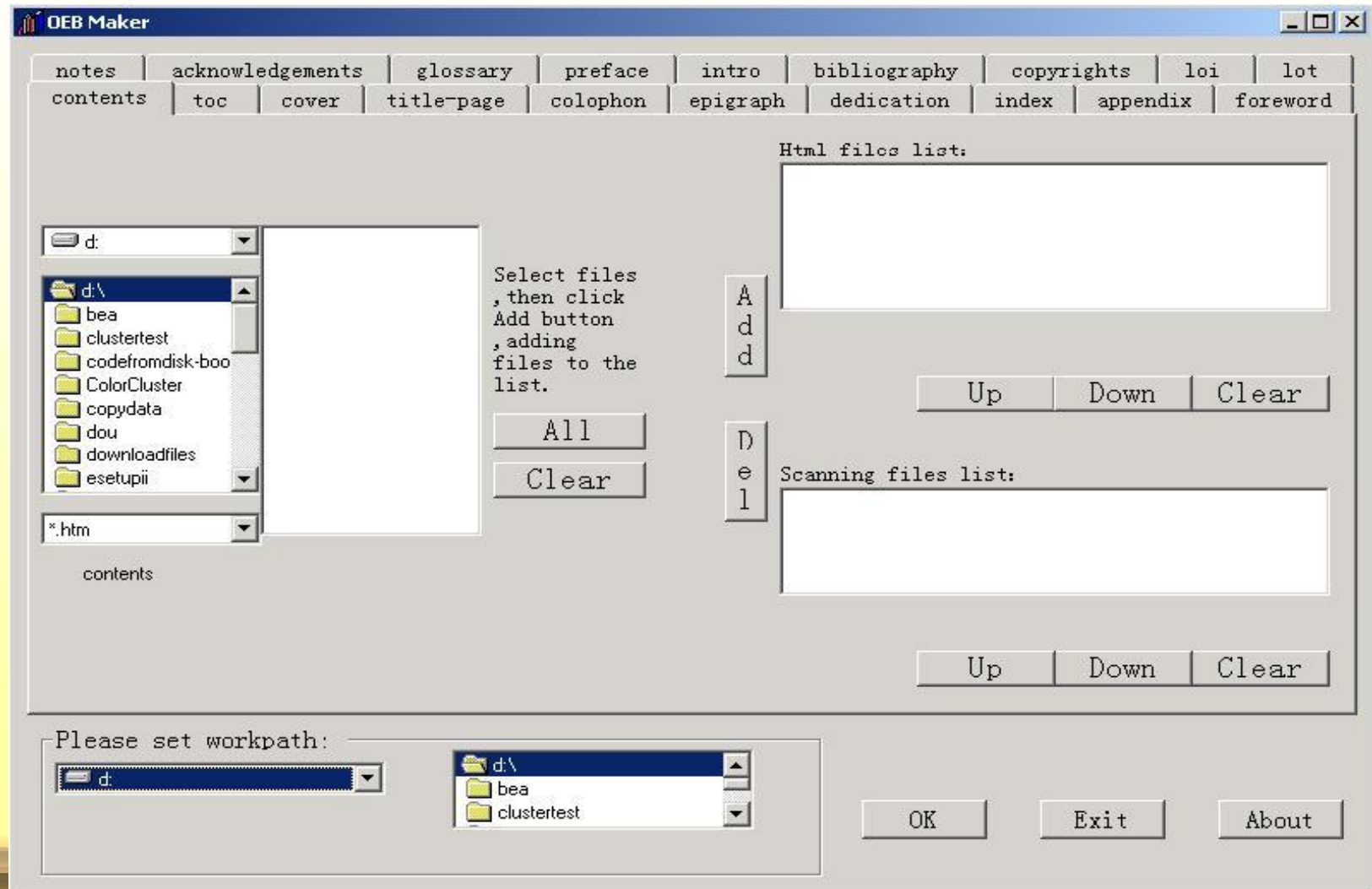
- Total: 100TB
- Two copies in China

System Architecture of ULIB @ China

E-Books

Book Scanning System

Books, papers

E-Book Quality Check Tools

Information Extraction Tools

E-Book Loading Tools

E-Book Content Management System

www.ulib.org.cn
Ulib @ China

Multimedia & Knowledge Service Systems

Knowmap

IQuery

Illustrator

……

E-Book Scanning Management System
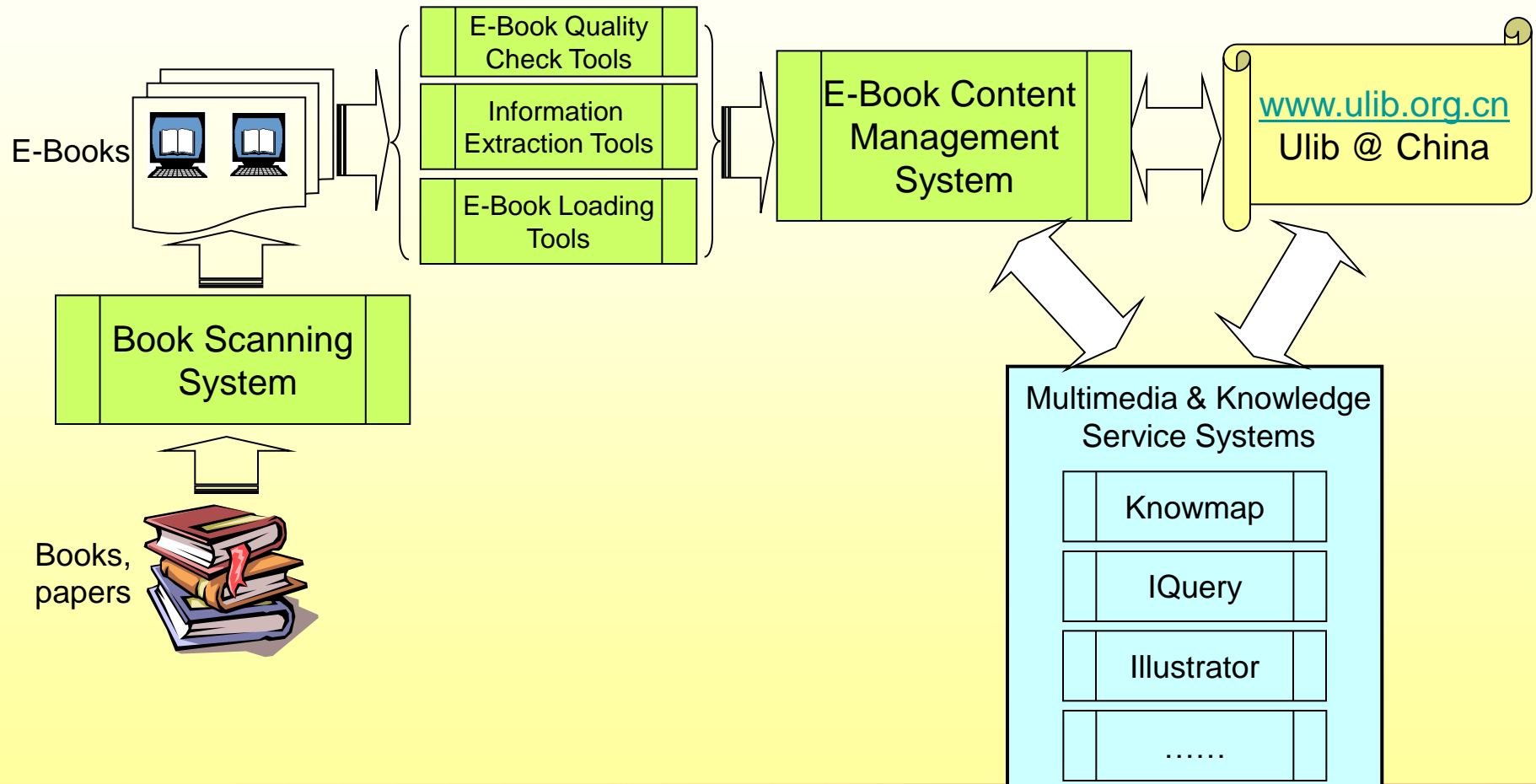
DCEditor: Creating Dublin Core Metadata
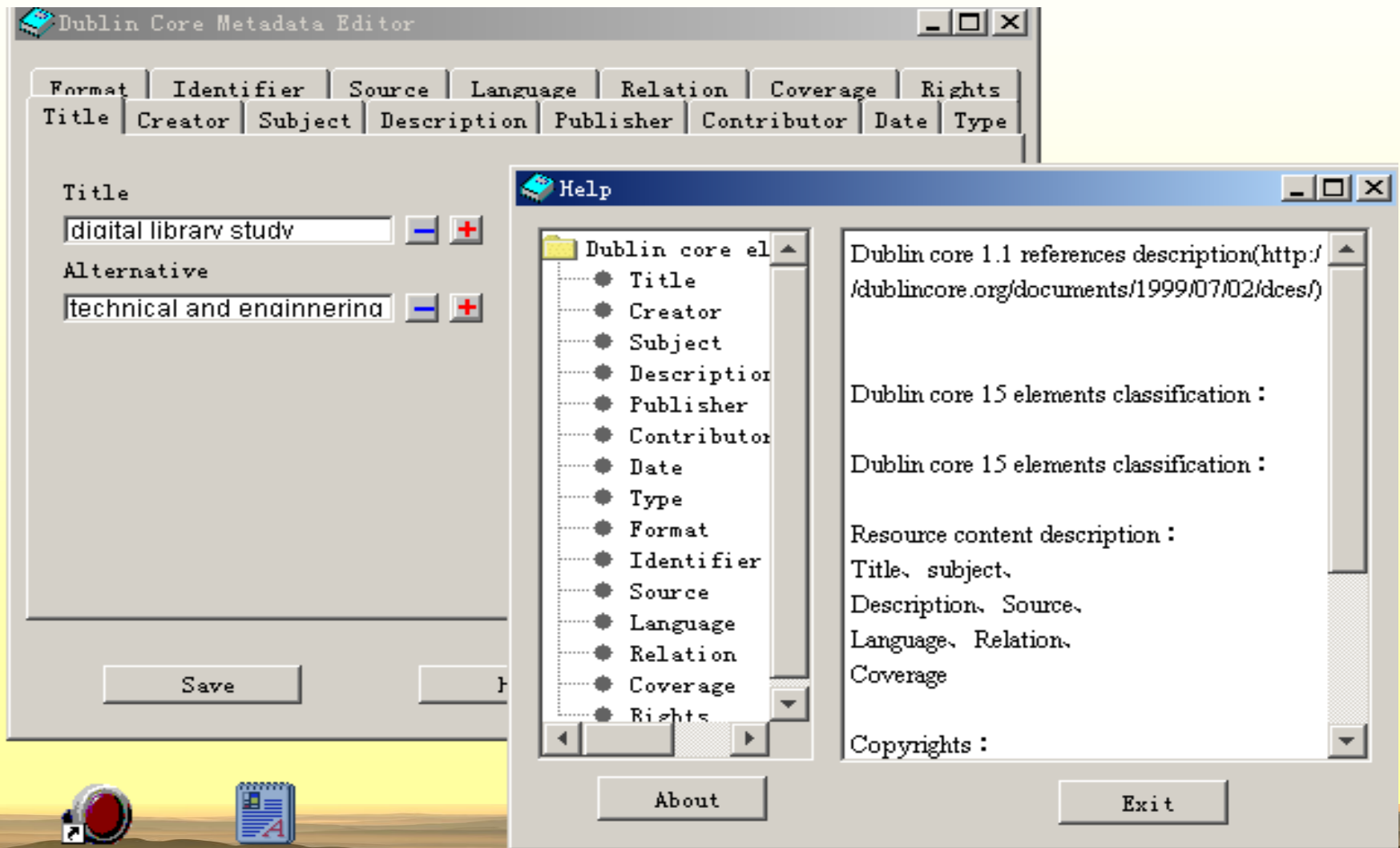
OEBMaker: E-Book Package with OEB

E-Book Quality Check Tool

Digital books entering archive......

根目录
├── 提交说明.txt
└── EBook
    ├── 图书1子目录
    │   ├── 1.txt
    │   ├── OTIFF
    │   ├── PTIFF
    │   ├── META
    │   ├── MARC
    │   ├── HTML
    │   ├── TOC
    │   ├── OEB
    │   └── oebbrowser.html
    ├── 图书2子目录
    └── 图书n子目录
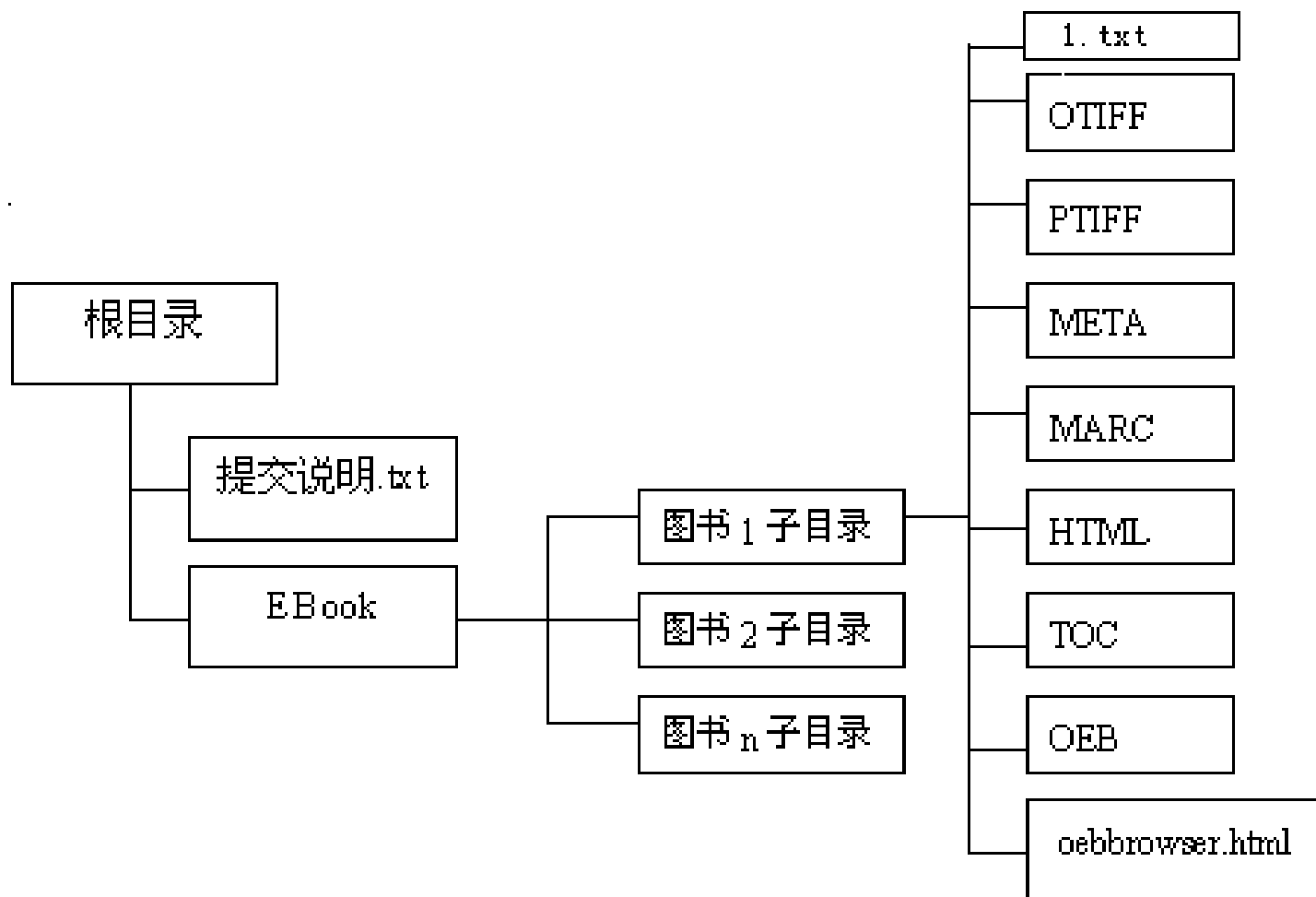
Futian Scanning Center focusing on English Book

Located at ShenZhen Futian Free trade area (Near Hongkang

Futian Scanning Center focusing on English Book

2005. 1. 22

China-US Million Book Digital Library

高等学校中英文图书数字化国际合作计划

英文数字图书质量控制要求

Quality Control Handbook for Million Book Project

"高等学校中英文图书数字化国际合作计划"北方技术中心

Northern Technical Center

中国科学院研究生院网络多媒体研究中心

for the China-US Million Book Digital Library Project

2003 年 11 月

Quality Control in Beijing

- Data QC
  - 16 persons
  - 200 titles per days
- Data backup and System maintain
  - 3 engineers
  - 650 titles per day
  - Storage, Web site
  - Data exchange

English Book Scanning

- **Shenzhen Futian Scanning Center**
  - 95 workers
  - 60 PCs and Servers
  - Scanners: 6 PS7000, 5 AVA3, and platform scanners
  - 160 titles or 50 thousand pages per day

- **Beijing Fangshan Scanning Center**
  - 60 workers
  - 30 PCs and Servers
  - Ten scanners
  - 100 titles or 30 thousand pages per day

书籍信息
　元数据信息
　　Title:World politics and interna
　　Creator:Bergsten, Fred
　　Subject:F11
　　Subject:International economi
　　Subject:International finance
　　Subject:World politics
　　Description:In recent years the
　　Publisher:The Brookings Institu
　　Contributor:Krause, Lawrence
　　Date:1975
　　Type:Text.Monograph
　　Format:text/html(.htm,.html)
　　Language:en
　　Rights:Copyright 1975 by broc
　　Date:2003-07-07
　目录浏览

　其他信息

C. FRED BERGSTEN AND LAWRENCE B. KRAUSE

Editors

World Politics and International Economics

Contributions by

C. Fred Bergsten, Robert O. Keohane, and Joseph S. Nye, Jr.
Robert Gilpin
Richard N. Cooper
Robert E. Baldwin and David A. Kay
Charles R. Frank, Jr., and Mary Baird
Robert O. Keohane and Van Doorn Ooms
Carlos F. Díaz-Alejandro

Example English book on China S

India Scanning Center

Book on Demand
in India Internet Café bar

English Books

- Before 2003
  - US side provided copyright-clear book list *(National Academy Press…),* China looked for book in libraries
  - Less than 1 thousand book
- US shipped books to China (2004~2006)
  - About 35 thousand
- HKU Library
  - About 35 thousand
- University City in Shenzhen and other Chinese Universities
  - About 85 thousand
- Totally, more the 150 thousand

English books in Digital (Up to Oct., 2006)

Books sources	Titles	Pages	Colour Pages
CMU (Carnegie Mellon Univ.)	24,537	7,645,950	342,402
NAL (National Agriculture L.)	2,083	624,900	28,405
HKU (Hongkong Univ.)	32,190	6,672,272	
UCSZ (Univ. City of Shenzhen)	~85,000		

Amount of Digital Books up to 2007

- China : more than 1 million
  - About 50% is copyright-clear and public available
  - Chinese portal @ http://www.cadal.zju.edu.cn
  - English books portal @ http://www.ulib.org.cn
- India: about 0.3 million
- Egypt: < 0.1 million

ULib Portal

The Universal Digital Library

Million Book Collection

Enter any word in the book title. e.g. Operas

Advanced Search | Browse the Collections

Mirror Sites China North   China South   India

For viewing UDL books you will need either **DjVu plugin** and/or Tiff plugin (**Windows** ,**Linux** , **MAC**)

We are currently performing system maintenance at our partner site at IIIT, Hyderabad, India. Some of the books hosted at this site may not be accessible online during this maintenance. Please try back in a few hours.

The UDL website is experiencing performance issues due to high traffic.We are also discovering errors in the metadata for the books. Please bear with us while we work to resolve these issues over the coming weeks and months. Click here for more details on our plans.

©Copyright 2001-2008, Universal Digital Library at Carnegie Mellon University & Partners

http://www.ulib.org hosted by CMU

ULib China Site

The Universal Library

China Site

Hosted by Graduated School of the Chinese Academy of Sciences

Book IQuery Illustrator Knowmap

Home

Browse

Experimental Systems ▾

MSearch *cnc*

MSearch *wiki*

iULib 1.0

IQuery

Illustrator

Knowmap

About

Million Book Project
Memorabilia
Goals
Download
Partners
Statistic
CopyRt Disclaimer

Rare Collections

Sikuquanshu

Recent Update

Collections

- CMU-SCS
- National Publishing House, USA
- Satish Dhawan's Library

Website Statistics

Since Dec 19th, 2007, we have:
Loading Count Error.

Search History

The Universal Digital Library

Million Book Collection at China Mirror Site

[SEARCH]

Enter any word in the Book Title, Eg:Operas

Search in 1,531,335 digital books!

Browse the Collections   Advanced Search

Million Book Search: Experimental Systems

iULib 1.0

iULib is the vision we proposed about where and how Wikiepdia and UDL can meet. In the first phase, iULib is a system which directly apply the Wiki architecture and software in UDL to upgrade the digital library as an interactive platform that facilitates community and collaboration.

Try out books in iULib 1.0

The Waste Land As Grail Romance: Eliot's Use Of The Medieval Grail Legends

Toward The Endless Frontierhistory Of The Committee On Science And Technology, 1959-79

IQuery 2.0 BETA

Phase II

- Apr04,2010 China government funds 150 million RMB to support to scan another 1.5 million books in next three years

- Partners
  - 15 Universities: Zhejiang U., Peking U., Tsinghua U….
  - Chinese Academy of Sciences
  - Shenzhen Scanning center

Ongoing New 1.5 Million......

http://www.cadal.net

Content

- Million Book Project
- Book Retrieval
  - Multilingual Access
  - Knowledge Organization and Retrieval
  - Illustration Retrieval
- Multimedia
  - Text-Image Retrieval
  - Video Annotating
  - Navigating Chinese Scrolls Fine Art

©北京大学黄铁军2010

Digital Library:
Universal Access to Human Knowledge

- Million Book Library:
  - 1% of books throughout recorded history worldwide
  - 100 times of books that every one can read throughout his life

- One million digitized books:
  - NOT to be the ultimate goal of the project
  - BUT a first step towards universal access to human knowledge.

Resources and Technologies

All Books Worldwide （100M ）

Other Resources

India 1M

CADAL 1M

Other(Egypt…)

Select by librarians

10 Thousand

Internet

Knowledge service platform

©北京大学黄铁军2010

Open the door of knowledge

- Cross-lingual access
  - Million Book Project integrates Chinese, English and other languages books
  - People need to directly conduct multi-linguistics retrieval for required information
- Efficient knowledge organization and services
  - To realize the mapping from one million and even more books in a library to ten thousand books for each reader
- Multimedia understanding
  - The human knowledge is not explicitly saved in the digital library, but is implicitly expressed in different media
  - Retrieval millions illustrations in digital book

Content

- Million Book Project
- Book Retrieval
  - Multilingual Access
  - Knowledge Organization and Retrieval
  - Illustration Retrieval
- Multimedia
  - Text-Image Retrieval
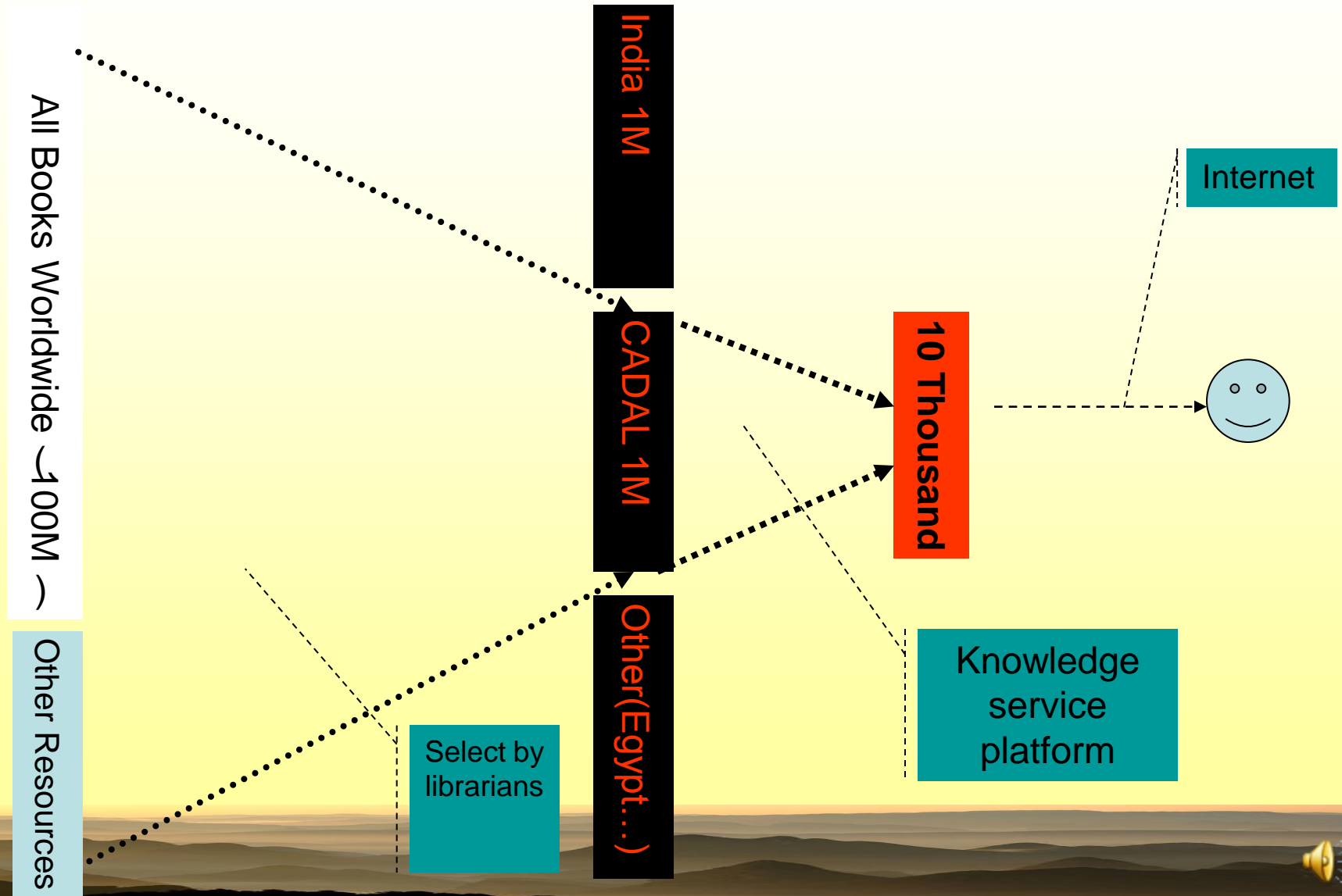 - Video Annotating
  - Navigating Chinese Scrolls Fine Art

Multilingual Access

- Access information in own language
  - Chinese people using Chinese to access information in any language
  - English people can understand what are in the Chinese library
- Our Work
  - Selective translation
  - Partial translation of ancient Chinese

Selective Translation

- ## Title translation
  - Disambiguate words and identify shortened form in the title by the whole text
  - EBMT for the whole title

- ## Important sentence extraction
  - According to titles or keywords

- ## Sentence compression
  - Basic approach

- ## Translation of selected sentences

Partial Translation of Ancient Chinese

- Identify named entities in Ancient Chinese
  - Combining statistical and rule-based means
- Translation of named entities
  - Statistical machine translation
  - Transliteration
- Translation of Ancient Poems
  - Statistical MT trained on 300 Tang Chinese-English Poems by Prof Yuancong Xu
  - Generations of phrases rather than whole sentences
  - Translation selection by rhyming words

Content

- **Million Book Project**
- **Book Retrieval**
  - **Multilingual Access**
  - **Knowledge Organization and Retrieval**
  - **Illustration Retrieval**
- **Multimedia**
  - **Text-Image Retrieval**
  - **Video Annotating**
  - **Navigating Chinese Scrolls Fine Art**

Traditional Solution(1): Metadata based Search

- Advantages
  - Simple implementation
  - High retrieval precision
- Disadvantages
  - The granularity of retrieval results is too big: a whole book as a retrieval unit.
  - It is too tiresome for a user to skim through the whole book to locate the required items.

Average 300 pages per boo

Traditional Solution(2): Full-text based Search

- Advantages
  - Directly locate the required items in books
  - High retrieval recall
- Disadvantages
  - The granularity of retrieval results is too small: all matched words in e-books as retrieval units.
  - Due to the overwhelming abundance of retrieval results, some kind of grouping .
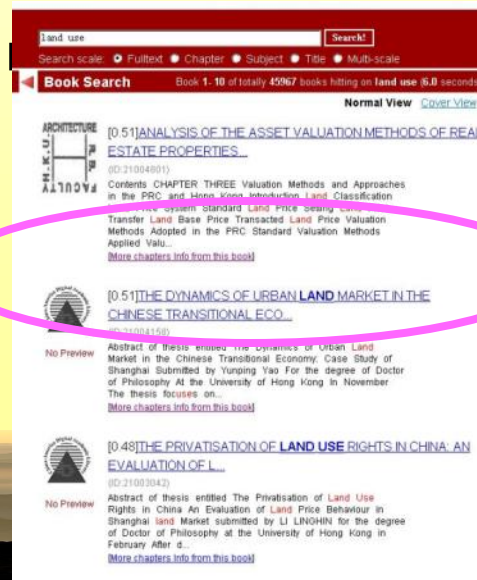
The Multi-granularity Model for E-Books

Facet info

BookNet →

ChapterNet →

TermNet →

Network among topically related books

Network among topically related chapters

Semantic network among terms

Legend

- 🔴 Book node
- 🔵 Chapter node
- ⚫ Term node
- ☐ Property
- —— Association
- - - - Corresponding
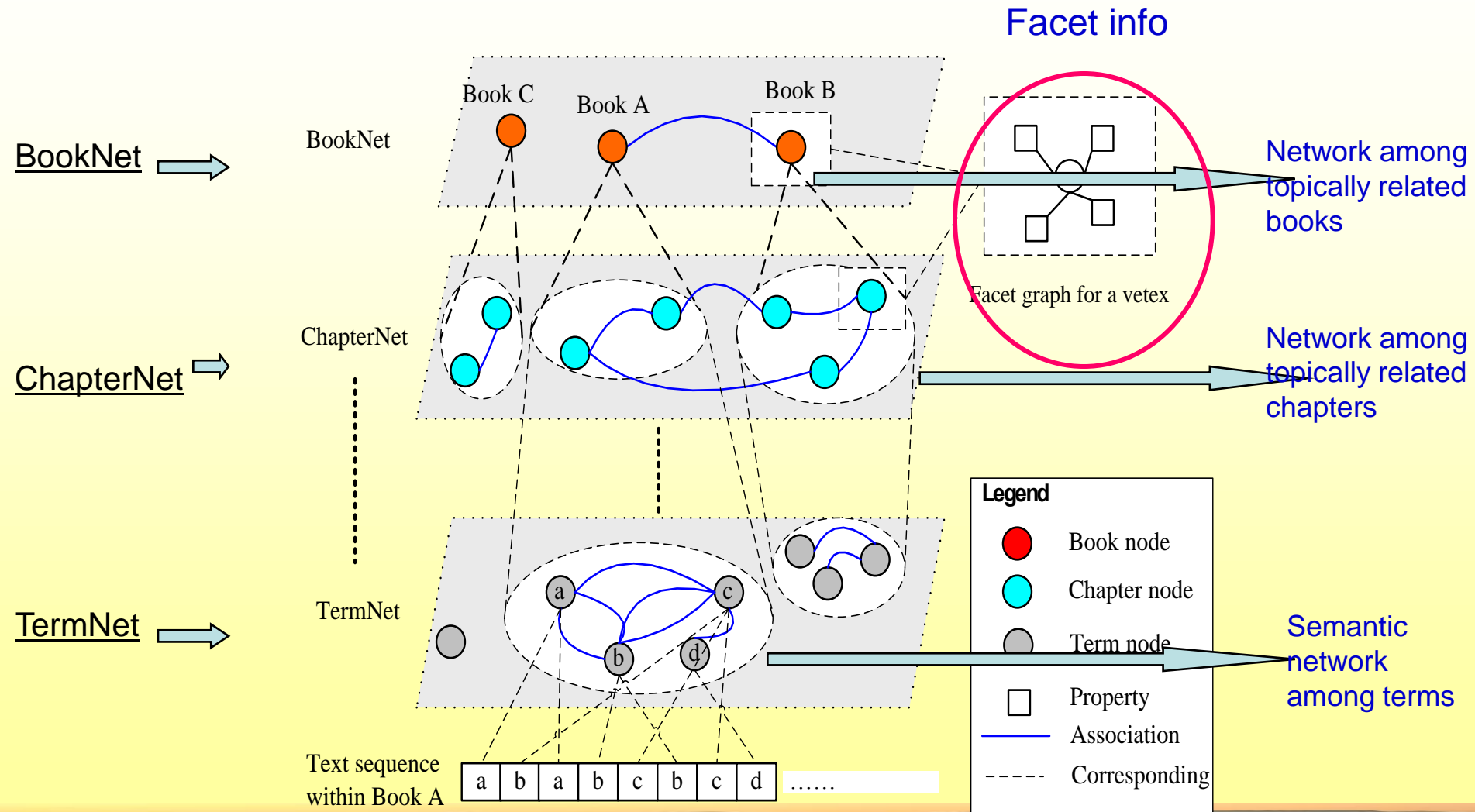
Context-based Semantic Modeling and Analysis

- What is "Context"
  - Definition: a collection of (external or internal) related objects or auxiliary information that make the semantics of that object unique and comprehensible.
  - Usage: to improve learning by exploiting context explicitly or implicitly.
- Models and Algorithms
  1. General model: Contextual dependency network models for relational data
  2. For key phrases: Key phrase extraction using semantic network structural analysis on small-world networks
  3. For e-books: Multi-granularity context modeling
  4. For illustrations: Cross-modal semantic analysis

Contextual dependency network models

- ## Key Ideas [TKDE'06][JIIS'06]

  1. To extend dependency networks for modeling relational data.

  2. To propose linkage semantic kernel to calculate contextual dependency in the link structure.

  3. To use the Gibbs inference framework for collective classification.
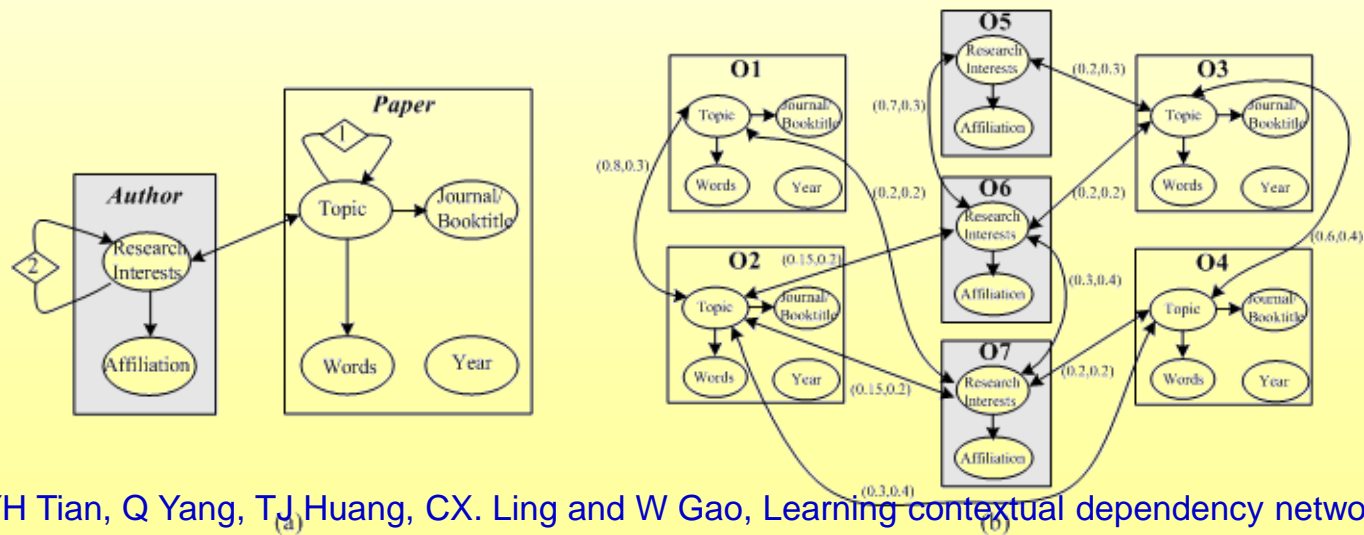
[TKDE'06] YH Tian, Q Yang, TJ Huang, CX. Ling and W Gao, Learning contextual dependency network models for link-based classification. IEEE Transactions on Knowledge and Data Engineering, 18(11), Nov 2006, 1482-1496.
[JIIS'06]YH Tian, TJ Huang, W Gao. Latent Linkage Semantic Kernels for Collective Classification of Link Data, Journal of Intelligent Information Systems, 26(3), May 2006, 269-301.

System DEMO

System Architecture

E-Books → Structure Analysis → Units at different granularities (Toc.xml) → Multi-level Info Extraction (Metadata Extracter, Key-phase Extracter, TFIDF Extracter) (DC / MARC) → Multi-granularity information → Model Builder (Similarity Evaluator, BookNet Builder, ChapterNet Builder, TermNet Builder) → MMKN

Book Layer, Chapter Layer, Term Layer
Book C, Book A, Book B

BookNet, ChapterNet, TermNet
Facet graph for a vetex
Text sequence within Book A

Backend

Front-end

Retrieval Model
- Multi-granularity Ranking
- Facet Grouping
- Information Visualizer

System Interface

Query expansion

■By exploiting the TermNet (e.g., "water supply", "water resource in USA" and so on for the query keyword "water").

Facet grouping

■Clustered into facets (e.g., people, time or place) according to different values of topic-related fields and subfields

Multi-granularity ranking

■To exploit multi-granularity information to improve the ranking of retrieval results

Information visualization

■To provide powerful capabilities for query refinement, navigation of related topical space, result analysis.

Subject Facets ◀ | Book Search | Book 1- 10 of totally 38203 books hitting on **Europe** (0.078 seconds) | Expanded ReSearch

Normal View   Cover View

🔍 Europe
- ⊞ 📁 Composite subject
- ⊞ 📁 General subfield
- ⊞ 📁 People
- ⊟ 📁 Time
  - 📄 1945 (65)
  - 📄 1918 (48)
  - 📄 20th century (38)
  - 📄 1989 (35)
  - 📄 1985 (14)
  - 📄 1789 (12)
  - 📄 1990 (12)
  - 📄 1981 (9)
  - 📄 1871 (7)
  - 📄 1975 (6)
  - 📄 20e si 768 ecle (6)
  - 📄 1979 (4)
  - 📄 1976 (3)
  - 📄 1984 (3)
  - 📄 Revolution 1789 (3)
  - 📄 1648 (2)
  - 📄 1769 (2)
  - 📄 1917 (2)
  - 📄 1919 (2)
  - 📄 1971 (2)
- ⊟ 📁 Place
  - 📄 United States (198)
  - 📄 Europe Eastern (81)
  - 📄 Soviet Union (57)
  - 📄 Great Britain (37)
  - 📄 European Union countries (30)
  - 📄 769 Etats (28)
  - 📄 URSS (17)
  - 📄 European Economic Communi (16)
  - 📄 EE UU (11)
  - 📄 China (10)

[0.41]Reshaping Europe: Strategies for a Post-Cold War Europe
(ID:33302143)

Subjects:@Europe|Foreign relations*United States; @United States|Foreign relations*Europe; @Europ...

Reshaping Europe: Strategies for Post-Cold War Europe Edited by Kim R. Holmes and Jay P. Kosminsky Reshaping Europe: Strategies for Post-Cold War Europe Edited by Kim R. Holmes and Jay P. Kosminsky ...
[More chapters Info from this book]

[0.41]REDEFINING EUROPE(ID:33322646)

No Preview

Subjects:@Internationale betrekkingen+gtt; @Europe|Politics and government^; @4+gtt; @Politics; ...

REDEFINING EUROPE New Patterns of Conflict and Cooperation Edited by Hugh Miall PINTER PUBLISHERS London and New York THE ROYAL INSTITUTE OF FI INTERNATIONAL AFFAIRS Hi Distributed in the United States and Canad...
[More chapters Info from this book]

[0.41]THE SHAPE OF THE NEW EUROPE(ID:33322562)

No Preview

Subjects:@European federation; @International economic relations; @Integraci 769 on europea; @Re...

THE SHAPE OF THE NEW EUROPE Edited by Gregory F. Treverton ON FOREIGN RELATIONS PRESS NEW YORK THE SHAPE OF THE NEW EUROPE Edited by Gregory F. Treverton ON FOREIGN RELATIONS PRESS NEW YORK ...
[More chapters Info from this book]

[0.41]Voice of America, Broadcast Schedule for languages other than Eng...
(ID:31043818)

Estonia:Eesti Radio Program etc Krasnoyarsk: Odessa: suSERBIAN to Europe see also CROATIAN Bajina Basta: Belgrade: Pozarevac: Zenica: SLOVAK to Europe see also CZECH Michalovce: Banska Bistrica: Trnava: Zvolen:

Expanded ReSearch

[1.0] United States
[0.47] 7 150
[0.3] Soviet Union
[0.28] Great Britain
[0.23] 769 Etats-Unis
[0.22] North Atlantic Treaty Organization
[0.2] World politics
[0.18] National security
[0.17] 4 935
[0.16] 4 240

[0.15] Buitenlandse politiek
[0.14] China
[0.14] European Union countries
[0.14] 4 200
[0.13] International economic relations
[0.13] International relations
[0.12] Asia
[0.12] European Union
[0.11] Japan
[0.11] Globalization

[0.11] Europe Eastern
[0.11] URSS
[0.1] Europese Unie
[0.1] Developing countries
[0.09] Democracy
[0.09] India
[0.09] NAVO
[0.09] Germany
[0.09] World War 1914-1918
[0.09] 5 310

[0.09] France
[0.09] EE UU
[0.09] Economische politiek

Multi-granularity Ranking

- Problem Formulation
  - To exploit multi-granularity information to improve the ranking of retrieval results
  - Example: For a query "neural network", a book with several relevant chapters > a book with only one relevant chapter, even with the same occurrence scores for "neural network".

- Key idea

Initial book relevance

Initial chapter relevance

Book	Relevance
A	0.8
B	0.6
C	0.55

Initial relevance scores for e-books

Chapter	Relevance
A.2	0.5
B.2	0.3
B.3	0.4
C.1	0.6
C.4	0.9

Initial relevance scores for chapters
(Assume that each e-book has ten chapters)

(a) Initial ranking by using intra-granularity information

$$r^{(1)}(b_i \mid q) = \theta r^{(0)}(b_i \mid q) + (1-\theta) \sum_{c_s \in b_i} \varpi_s r^{(0)}(c_s \mid q)$$

Control factor of inter-granularity affect

Weights to measure the topical relevance of a chapter on the whole book

Updated book relevance

Rank	Book
1	A
2	C
3	B

(b) Re-ranking of e-books by taking into account the affects of chapter relevance scores on e-book ranking

Rank	Chapter
1	C.4
2	A.2
3	C.1
4	B.3
5	B.2

(c) Re-ranking of chapters by taking into account the affects of book relevance scores on chapter ranking

Experimental Results

- Baselines
  - MDSearch: Metadata based search
  - FTSearch: Fulltext based search

Query	FTSearch		IQuery			MDSearch	
	s	N	s(c)	s(b)	N	s	N
control	3.0 ± 1.9	10	7.2 ± 2.7	5.6 ± 2.3	10	7 ± 2	10
mathematics	4.8 ± 3.4	10	N/A	N/A	0	6 ± 3	9
beauty	4.8 ± 2.5	10	5.2 ± 2.1	4.6 ± 1.6	10	N/A	0
education	6.3 ± 2.7	10	8.2 ± 1.8	6.3 ± 3.0	10	8 ± 2	4
multimedia	N/A	0	N/A	N/A	0	N/A	0
protocol	N/A	0	N/A	N/A	0	N/A	0
China[*]	N/A	0	6.0 ± 2.1	4.4 ± 0.8	5	10	1
culture	5.8 ± 2.7	10	5.4 ± 3.3	4.2 ± 2.6	10	3 ± 1	3
war[*]	2.0 ± 0	10	7.7 ± 2.5	6.0 ± 2.9	10	7 ± 3	10
health	5.2 ± 3.3	10	5.8 ± 3.7	6.4 ± 3.6	10	2	1
	± 1.4	10	8.0 ± 2.0	5.3 ± 1.2	3	N/A	0
	± 2.5	10	6.8 ± 2.9	5.8 ± 2.4	10	4	1
	± 3.6	10	7.0 ± 3.7	6.0 ± 3.7	10	6 ± 6	2
symphony	4.4 ± 2.8	10	8.0 ± 0	8.0 ± 0	3	N/A	0
sculpture	7.0 ± 3.3	10	10.0 ± 0	10.0 ± 0	10	10	1

Note: s is short for Mic s@10; s(c) and s(b) stand for Mic s@10 when returned chapters and books respectively; N is short for NoR@10.

Legend:
- FTSearch
- IQuery with s(c)
- IQuery with s(b)
- MDSearch

Macro average s@n for all three systems

Content

- **Million Book Project**
- **Book Retrieval**
  - **Multilingual Access**
  - **Knowledge Organization and Retrieval**
  - **Illustration Retrieval**
- **Multimedia**
  - **Text-Image Retrieval**
  - **Video Annotating**
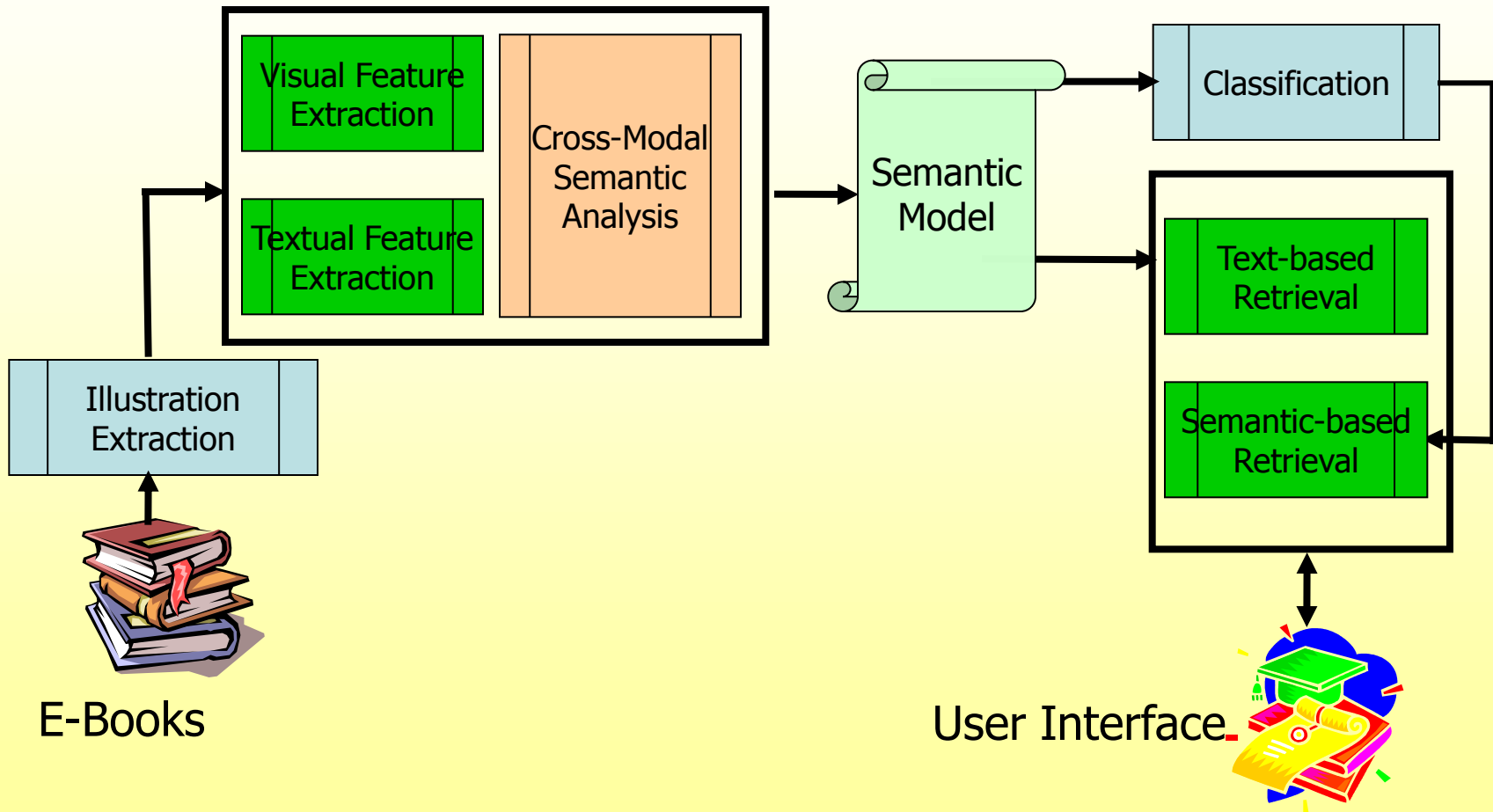 - **Navigating Chinese Scrolls Fine Art**

System Architecture

```
                    ┌──────────────────────────────────┐
                    │  ┌─────────────┐  ┌────────────┐  │        ┌──────────────┐
                    │  │ Visual      │  │            │  │        │ Classification│
                    │  │ Feature     │  │ Cross-Modal│  │        └──────────────┘
        ┌──────────┐│  │ Extraction  │  │ Semantic   │  │   Semantic
        │          ││  └─────────────┘  │ Analysis   │  │   Model    ┌──────────────┐
Illustration       ││  ┌─────────────┐  │            │  │            │ Text-based   │
Extraction         ││  │ Textual     │  │            │  │            │ Retrieval    │
        └──────────┘│  │ Feature     │  └────────────┘  │            ├──────────────┤
                    │  │ Extraction  │                  │            │ Semantic-based│
                    │  └─────────────┘                  │            │ Retrieval    │
                    └──────────────────────────────────┘            └──────────────┘
```

Visual Feature Extraction

Textual Feature Extraction

Cross-Modal Semantic Analysis

Illustration Extraction

Semantic Model

Classification

Text-based Retrieval

Semantic-based Retrieval

E-Books

User Interface

Illustration Features

- Textual features – Words Surrounding

HTML Tags	ATL field of IMG	TITLE	H1	H2	H3	H4	H5	H6	B	I	<No Tag>
Weights	6.00	5.00	4.00	3.60	3.35	2.40	2.30	2.20	3.00	2.70	1.00

- Visual features

  – "Image" Illustration:

  Moment Invariants

  – "Graphical" Illustration:
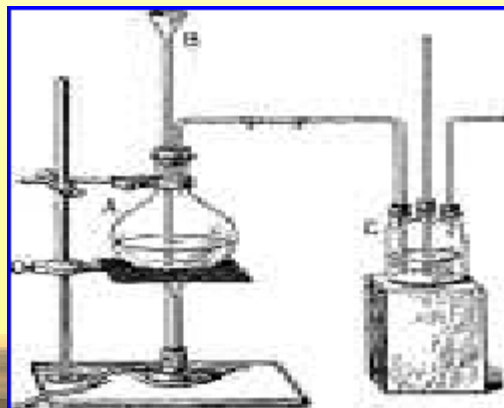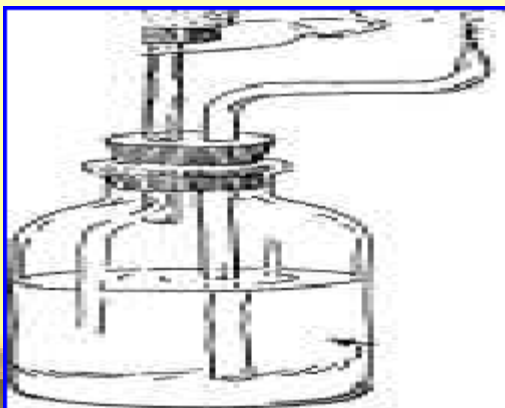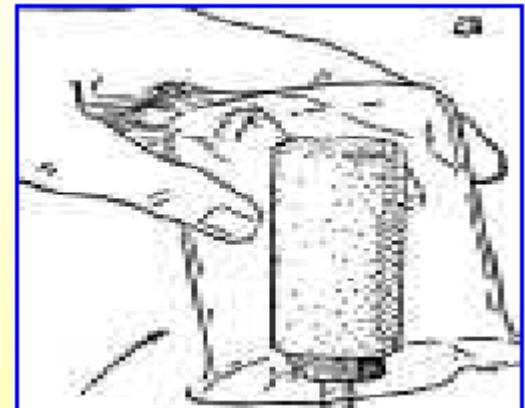
  Get primary shapes (e.g., Line and Arc) sequence,
  Transform to a sub-sequence that keeps most information

Retrieval Example

Input: bottle
Results from about 220,000 illustrations
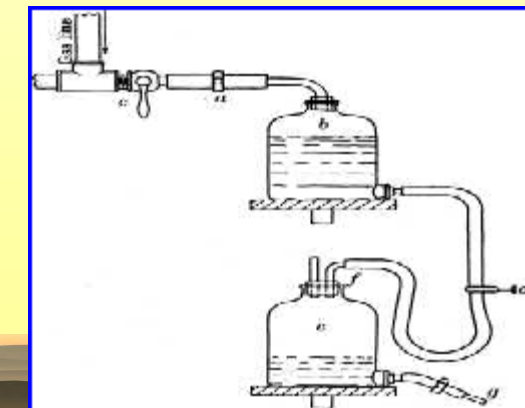
Diversifying the Retrieval Results via Topic Coverage Graph

- Topic Coverage Graph (TCG)
  - To represent the degree of mutual topic coverage among illustrations
  - Topic coverage is defined by

$$m_{jk} = sim(i_j, i_k) = \begin{cases} \dfrac{|W_k \cap W_j|}{|W_k|} & k \neq j \\ 0 & k = j \end{cases}$$

  Descriptive keywords for illustration i_j

- TRS-based Re-ranking

- Topic Richness Score (TRS)
  - To measure the importance of each illustration for topic coverage of retrieval results.
  - Random walk based calculation approach

$$\lambda = c\tilde{M}^T\lambda + \frac{1-c}{n}\vec{e}$$

  Normalized TCG matrix

  TRS vector

  Dumping factor

  Unit vector

Image Set → Topic Coverage Graph → Topic Richness Score

Query → Relevant Results

Output / Re-rank (+)

61
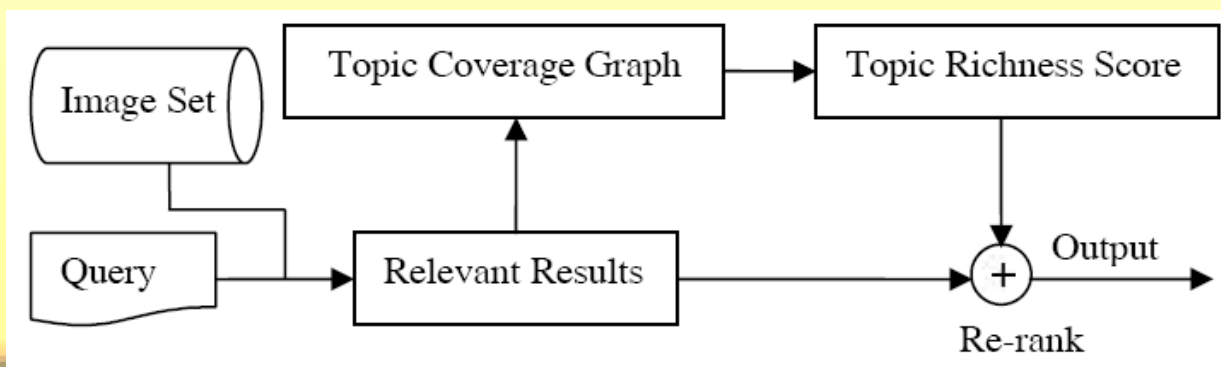
Example of Retrieval Topic Coverage Diversifying

- Before

For query "Forest"
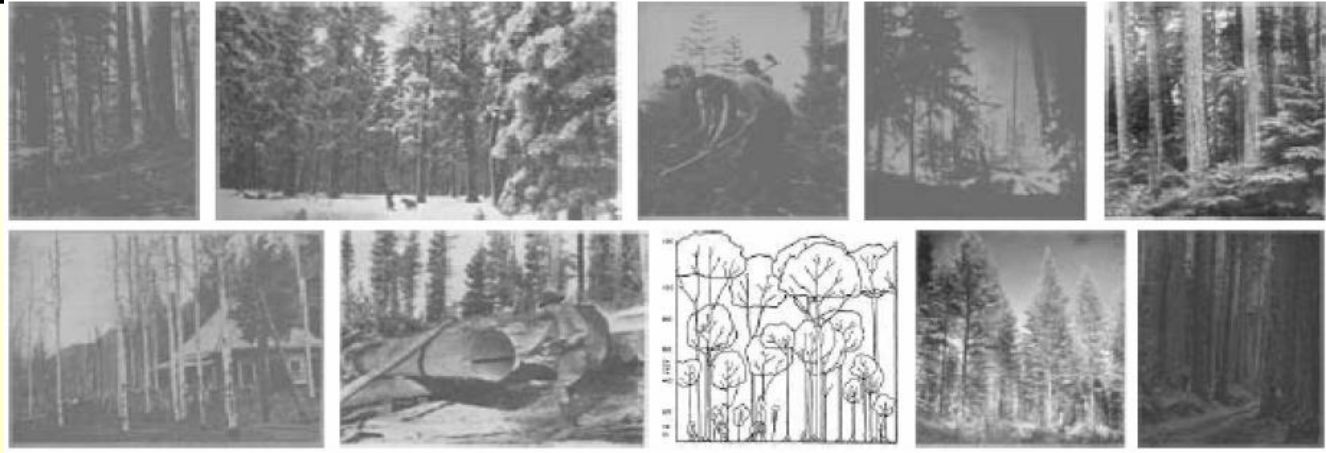
Most results are about the forest panorama.

- After TRS-based Re-ranking

For query "Forest"

A   B   C   D   E
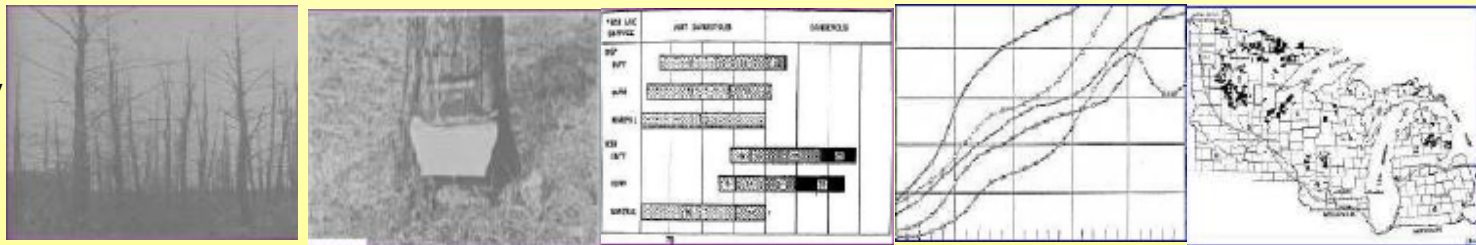
A: Forest panorama, B: Close-up of wood, C: Distribution of seldom woods, D: Forest precipitation graph, E: Distribution of forest area in USA.

62

Experiments

- Data collection
  - About 0.5 Million illustrations extracted from English e-books

- Query
  - 20 query words, respectively fed in Illustrator and TBR (Text-based Retrieval system)

- Evaluation Measures
  - Precision of Top-10 results
  - Topic diversity

System Demo

Content

- Million Book Project
- Book Retrieval
  - Multilingual Access
  - Knowledge Organization and Retrieval
  - Illustration Retrieval
- Multimedia
  - Text-Image Retrieval
  - Video Annotating
  - Navigating Chinese Scrolls Fine Art

Wikipedia as a knowledge base

- Lessons from these practices
  - It's helpful to build a KB for empowering the DL services.
  - Existing KBs are constructed NOT from users' contributions.
  - TermNet is not enough for query expansion in book retrieval
- Three usages of Wikipedia in DLs [Frumkin, 2005]
  - As a tool to empower interactive finding aids - AskWiki
  - As a content management tool – iUlib 1.0
  - As a knowledge base tool – Wikitology, iUlib 2.0
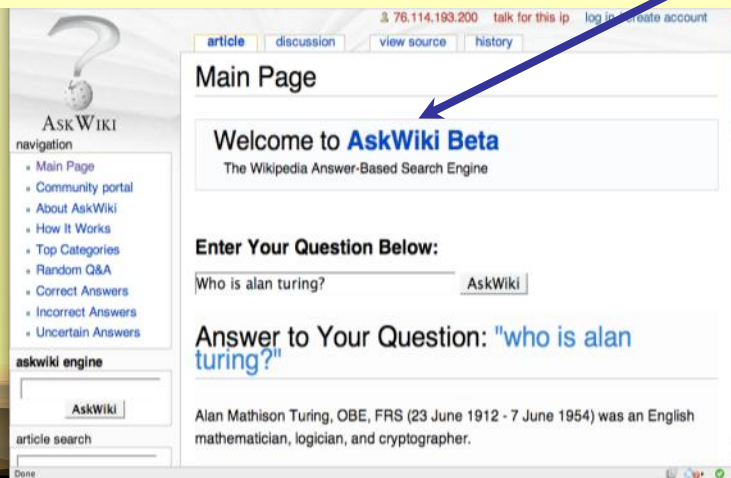
How to construct KB from Wikipedia?

- PF-IBF

    - PF-IBF: Path Frequency – inverse Backward link Frequency, proposed in [Nakayama,2007];

    - If there are many paths between two concepts, then they are possible to have strong correlation.
      → *pf* is calculated

$$pf(v_i, v_j) = \sum_{i=1}^{n} \frac{c_n}{f(n)}$$

c_n: the count of paths with length of n;
f(n): a monotonically increasing function enabling that a longer path length would lead to a smaller weight

    - *ibf*: A factor to balance the different link numbers

$$ibf(t) = \log(\frac{N}{bf(t) + \beta})$$

ibf(t): the number of backward links of concept t;
N: the total number
β: constant to prevent zero dividing

    - *Relatedness* between two concepts:

$$pfibf(v_i, v_j) = pf(v_i, v_j) \cdot ibf(v_j)$$

Examples of KBs

Query term	TF-IDF	Co-occurrence	PF-IBF	TermNet
Data Mining	Relational data mining, Data stream mining, Java Data Mining, Web data mining, Oracle Data Mining, Weather Data Mining	Data, Knowledge discovery, Data warehouse, Enterprise resource planning, Java Data-Mining	Human rights, Reliable sources, Open Directory Project, Information, Database, Education, Data warehouse	MINERA, Database management, Database searching, Temporal databases, Computer algorithms
Thesaurus	Crisp Thesaurus, RCA Thesaurus, Open Thesaurus, Thesaurus.com	Historical Thesaurus of English, Roget's Thesaurus, Terminology, Synonyms	Wiktionary, Dictionary, Oxford English Dictionary, Controlled vocabulary	N/A
Land Use	Nonconforming use (land use), Special use permit (land use), Variance (land use), Land use planning, Land use forecasting, Locally unwanted land use, Land Use and Development Act	Variance (land use), Zoning, Deforestation, Wilderness, Salinization, Land use planning, Natural resources	diversity, Zoning, ral environment, estation, pogenic, Soil erosion	City planning, Town planning, Environmental policy, Regional planning, Transportation, Urban renewal, Agriculture
City Planning	Planned city, Garden city (city planning), List of planned cities, Master of City Planning, City planning effort	Regional planning, Satellite campus, Tract housing, Urban design, Cluster developm	nts, Zoning, strial park, Cemetery,	Urbanisme, Town planning, Cities and towns, Urban renewal, ning
Peking Univ.	University of Peking, Peking Medical University, High School of Peking University, Peking University alumni, Peking University faculty	Tsinghua Univers Fudan University, Zhejiang Universi Hong Kong Unive of Science and Technology		
Iraq War	Iraq War/Iraq Insurgency, Iraq War/Iraq Timeline, 2003_war_in_Iraq, 2003_Iraq_war, The Iraq War, 1991_Iraq_war	Iraq, Sectarian, Saddam Hussein, resistance movement		War on rld politics, Imperialism

•Link co-occurrence and PF-IBF, have an obvious advantage that they can discover latent correlations.
•PF-IBF is unstable sometimes.

©北京大学 2010

Experiments at ImageCLEF 2008

- ImageCLEF: the cross-language image retrieval track
- WikipediaMM Task:
  - 151,519 web images, 75 topics
  - To investigate effective retrieval approaches in the context of a large-scale and heterogeneous collection of Wikipedia images.
- Our solution
  - Query extension with KB constructed from Wikipedia
  - Cross-media re-ranking by combining text-based and content-based image retrieval
  - 1st rank in all submitted runs

Experiments at ImageCLEF 2008

- Query expansion with KB constructed from Wikipedia

Near Pages Selection： Index all titles of Wikipedia articles, and then retrieve the Wikipedia pages with TF-IDF model.
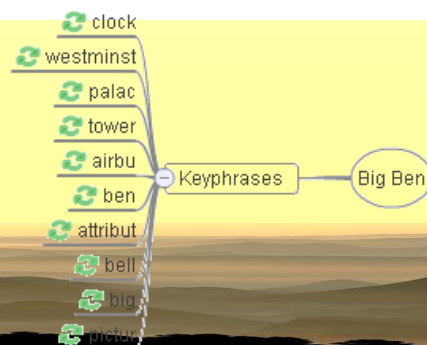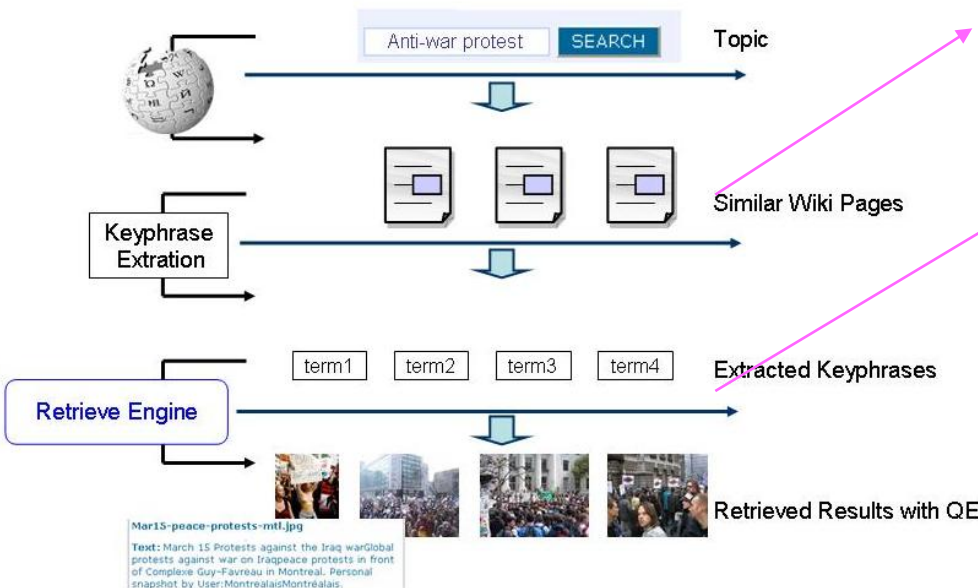
Page Keyphrase Extraction： Extract keyphrases with SWN(Small World Network) from selected pages.

$$K = \{(t_k, P(t_k))\}$$

Term Selection for Query Extension： Combine the importance and link frequency of keyphrases. (PF-IBF)

$$w_{t_k \in K} = P(t_k) \cdot ibf(t_k)$$

$$ibf = \log(\frac{N}{bf(t) + \beta})$$

Experiments at ImageCLEF 2008

- ## Query expansion with KB constructed from Wikipedia is effective!

	Participant	Run	Type	Modality	Feedback/Expansion	MAP	P@5	P@10	R-prec.	Bpref	Number of retrieved documents	Number of relevant retrieved documents
1	upeking	zzhou3	AUTO	TXT	QE	0.3444	0.5733	0.4760	0.3794	0.3610	18839	3406
2	cea	ceaTxtCon	AUTO	TXTCON	QE	0.2735	0.5467	0.4653	0.3225	0.2981	12577	2186
3	ualicante	IRnNoCamel	AUTO	TXT	NOFB	0.2700	0.4533	0.3893	0.3075	0.2728	47972	3497
4	cea	ceaTxt	AUTO	TXT	QE	0.2632	0.5200	0.4427	0.3080	0.2837	12577	2184
5	ualicante	IRnNoCamelLca	AUTO	TXT	FB	0.2614	0.4000	0.3587	0.2950	0.2577	74167	3624
6	ualicante	IRnNoCamelGeog	AUTO	TXT	QE	0.2605	0.4267	0.3640	0.3000	0.2646	48948	3423
7	ualicante	IRnConcSinCamLca	AUTO	TXTCON	FB	0.2593	0.3973	0.3493	0.3016	0.2617	73287	3538
8	ualicante	IRnConcSinCam	AUTO	TXTCON	NOFB	0.2587	0.4240	0.3627	0.2975	0.2631	55663	3505
9	ualicante	IRnNoCamelLcaGeog	AUTO	TXT	FBQE	0.2583	0.4000	0.3613	0.2922	0.2548	74167	3591
10	sztaki	bp_acad_textonly	AUTO	TXT	NOFB	0.2551	0.4400	0.3653	0.3020	0.2711	48847	3338

See http://imageclef.org/2008/wikimm-results

PKU at ImageCLEF 2008: Experiments with Query Extension Techniques

for Text-based and Content-based Image Retrieval

Zhi Zhou[1,2], Yonghong Tian[2], Yuanning Li[1,2], Ting Liu[1,2], Tiejun Huang[2], Wen Gao[2]

[1]The Institute of Digital Media, School of EE & CS, Peking University, Beijing 100871, China

[2]The Institute of Computing Technology, Chinese Academy of Science, Beijing 100080, China

{zzhou, ynli, tliu}@jdl.ac.cn, {yhtian,tjhuang,wgao}@pku.edu.cn

Image Search – "E"

http://www.ulib.org.cn/msearch-wiki/crossmedia.jsp

Image Search – "Eu"

Image Search – "Euro"

Image Search – "Europe"

Experimental System@WikipediaMM Task of ImageCLEF 2008 | 北京大学数字媒体研究所

Image Search – "Europe" next 10

Content

- **Million Book Project**
- **Book Retrieval**
  - **Multilingual Access**
  - **Knowledge Organization and Retrieval**
  - **Illustration Retrieval**
- **Multimedia**
  - **Text-Image Retrieval**
  - **Video Annotating**
  - **Navigating Chinese Scrolls Fine Art**

©北京大学黄铁军2010

Motivation(1)

– Semantic Gap

Server

Content based video annotation and indexing

Query by concept / lexicon

Video Corpus

Semantic Gap

User

– Concepts(lexicons) annotation from visual content

plane, clouds, sky   Fox, snow, wood

Motivation(2)

- ## Annotate over individual shots

Labeled training samples:

Building   Car   Human   Ship   Vegetation

time →

Car	1	1	1	1	1	1	1	1
Building	1	0	0	1	1	1	1	0
Sky	0	0	0	0	0	1	1	0
Street	1	0	0	1	0	1	1	0

2010-4-27

- Sequence multi-labeling
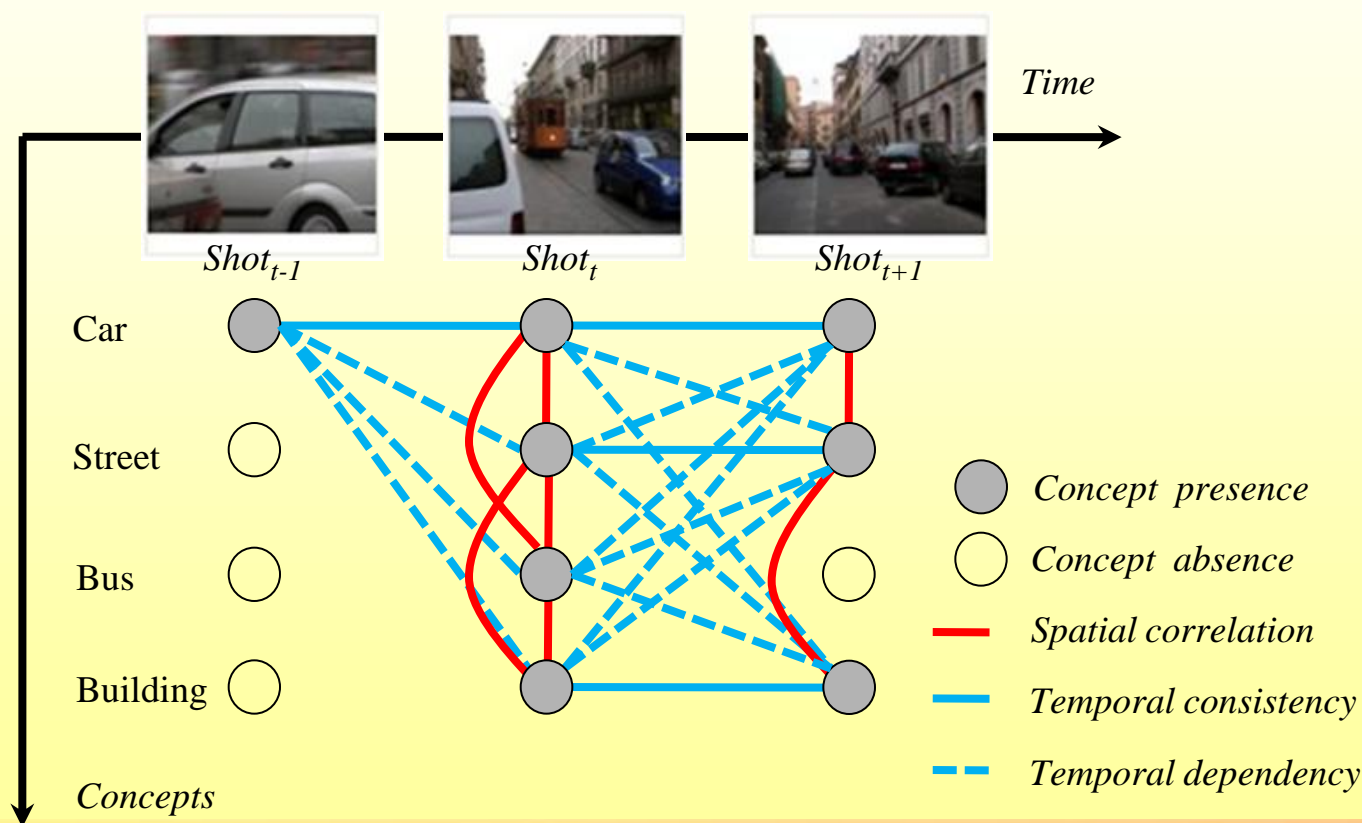
Temporal correlation of concepts

- The 1st, 2nd, and 3rd order temporal correlations

(a)   (b)   (c)

Calculated over the development dataset of TRECVID 2007

©北京大学黄铁军2010
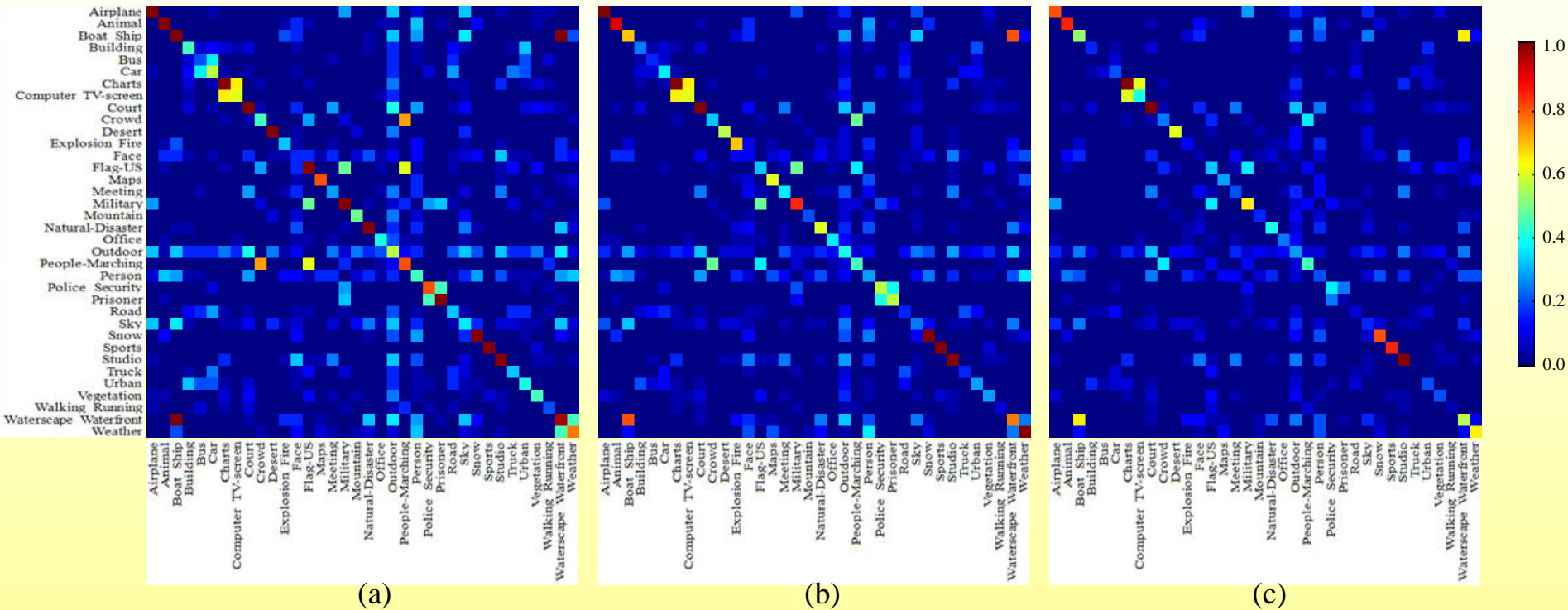
Three paragms for video annotation

(a) IML (b) IML-T (c) SML

Spatial correlation
Temporal consistency
Temporal dependency
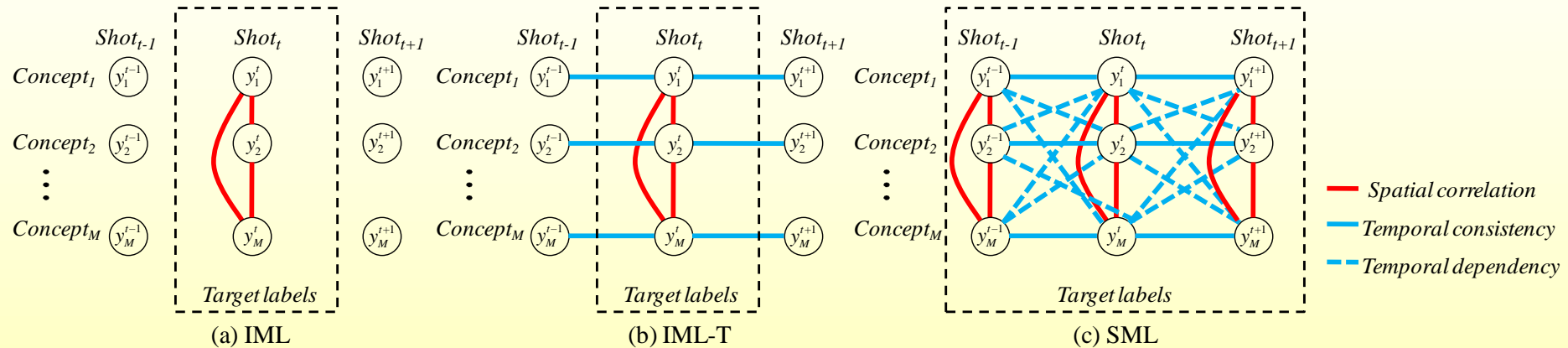
- Multi-labeling over individual shots (IML)
- Temporal refinement over IML (IML-T)
- Our proposed: Sequence Multi-labeling (SML)

©北京大学黄铁军2010

SVM$^{\text{SML}}$:A discriminative method for SML

– Joint kernel to model the similarity of shot sequences

$$K^{\text{joint}}\big((\mathbf{x},\mathbf{y}),(\bar{\mathbf{x}},\bar{\mathbf{y}})\big) = \big\langle \Phi(\mathbf{x},\mathbf{y}),\Phi(\bar{\mathbf{x}},\bar{\mathbf{y}})\big\rangle$$

$$= \sum_{t,\bar{t}}\sum_{w}\sum_{l=1}^{L}\beta_l^w K_l(\mathbf{x}^{t-w},\bar{\mathbf{x}}^{\bar{t}-w})\sum_{1\le p\le M}\Big[y_p^t = \bar{y}_p^{\bar{t}} = 1\Big]$$

Similarity in low-level feature

$$+ \sum_{t,\bar{t}}\sum_{w}\beta_{L+1}^w \sum_{1\le p\le M}\sum_{1\le q\le M}\Big[y_p^t = \bar{y}_p^{\bar{t}} = 1 \wedge y_q^{t-w} = \bar{y}_q^{\bar{t}-w} = 1\Big],$$

Spatial and temporal correlation

– Kernel based structure SVM

$$F(\bar{\mathbf{x}},\bar{\mathbf{y}}) = \big\langle \mathbf{w},\Phi(\bar{\mathbf{x}},\bar{\mathbf{y}})\big\rangle$$

$$= \sum_{i,\mathbf{y}\in\kappa/\mathbf{y}_i}\alpha_i(\mathbf{y})\Big[K^{\text{joint}}\big((\mathbf{x}_i,\mathbf{y}_i),(\bar{\mathbf{x}},\bar{\mathbf{y}})\big) - K^{\text{joint}}\big((\mathbf{x}_i,\mathbf{y}),(\bar{\mathbf{x}},\bar{\mathbf{y}})\big)\Big]$$

Framework of SML

A training set of sequence of shots

Test video shot sequence

Predicted multi-label sequence

$$\{...,(1,0,0,0,...)_{T-2},(1,1,1,1,...)_{T-1},(1,0,1,1,...)_{T}\}$$

T-2 T -1 T

Multi-modal feature extraction

SVM^{SML}

Joint kernel mapping

Overlapping features-label dependencies

Label temporal correlations

Label spatial correlations

Sequence multi-label classifier

Joint kernel optimization

Working set learner

Binary MRF solver

Shot ranking

Car

User query

Sequence of multi-label

Training path

Testing path

©北京大学黄铁军2010

Evaluation(1)

- Datasets
  - TRECVID'05 (170 hours news video)
  - TRECVID'07 (50 hours documentary video)
- State-of-the art features and basic kernels
  - Features
    - Dense-color-SIFT (DCSIFT) and dense-SIFT (DSIFT)
    - Self-similarity and pyramid histogram of orientated gradients (PHOG)
    - Gabor and co-occurrence texture
  - Kernels
    - Spatial Pyramid Kernel
    - Proximity Distribution Kernel
- Baseline Methods
  - Independent SVM
  - Correlative Multi-labeling (MSRA)
  - Temporal Correlative Multi-labeling (MSRA)

Evaluation(2)

- ## MAP 50.56% over TRECVID'05 dataset
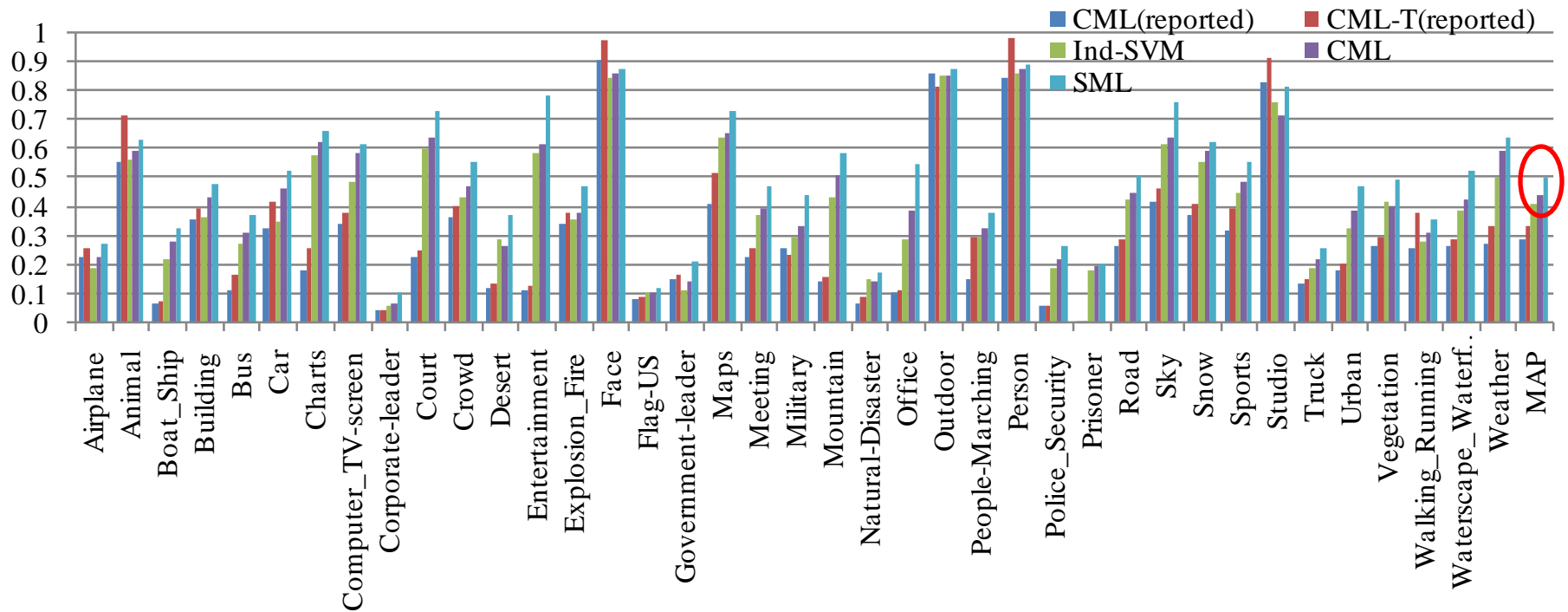
Evaluation(3)

- MAP 38.66% over TRECVID'07 dataset
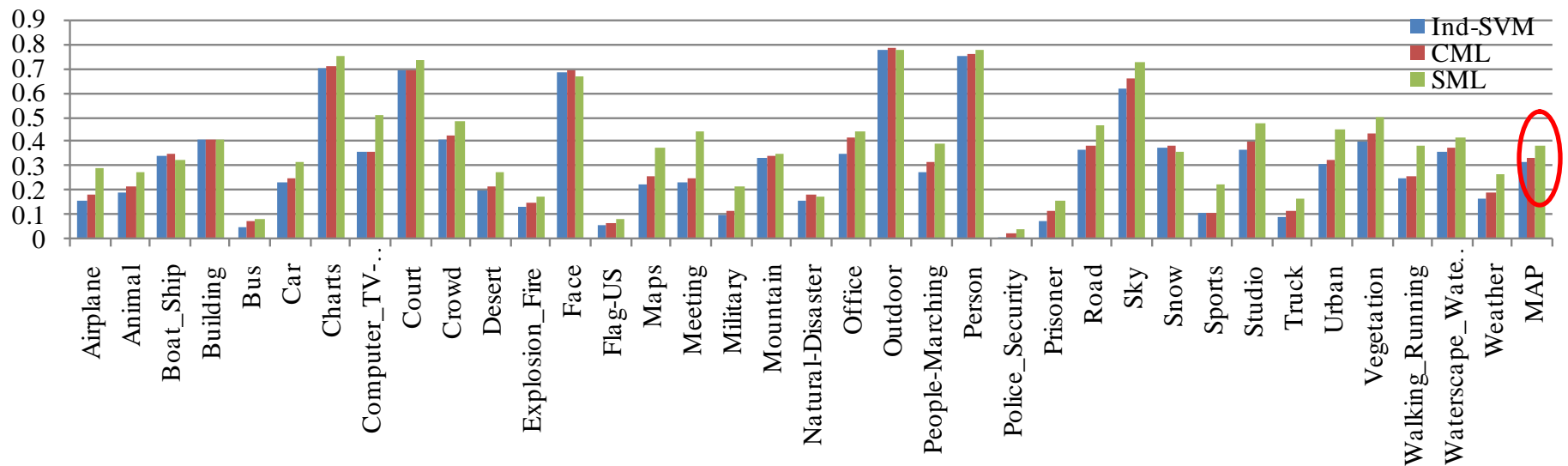
Content

- **Million Book Project**
- **Book Retrieval**
  - **Multilingual Access**
  - **Knowledge Organization and Retrieval**
  - **Illustration Retrieval**
- **Multimedia**
  - **Text-Image Retrieval**
  - **Video Annotating**
  - **Navigating Chinese Scrolls Fine Art**

Annotating Chinese Scrolls for Exhibition

- Collaboration partners of the project

  - IDM, Peking University

  - The Palace Museum of China

  - Microsoft Research Asia

- Motivations

  - Appreciation of arts

  - Exhibition of the artworks' details

  - Revelation of the people's life styles in the scrolls

  - Immersive user experience

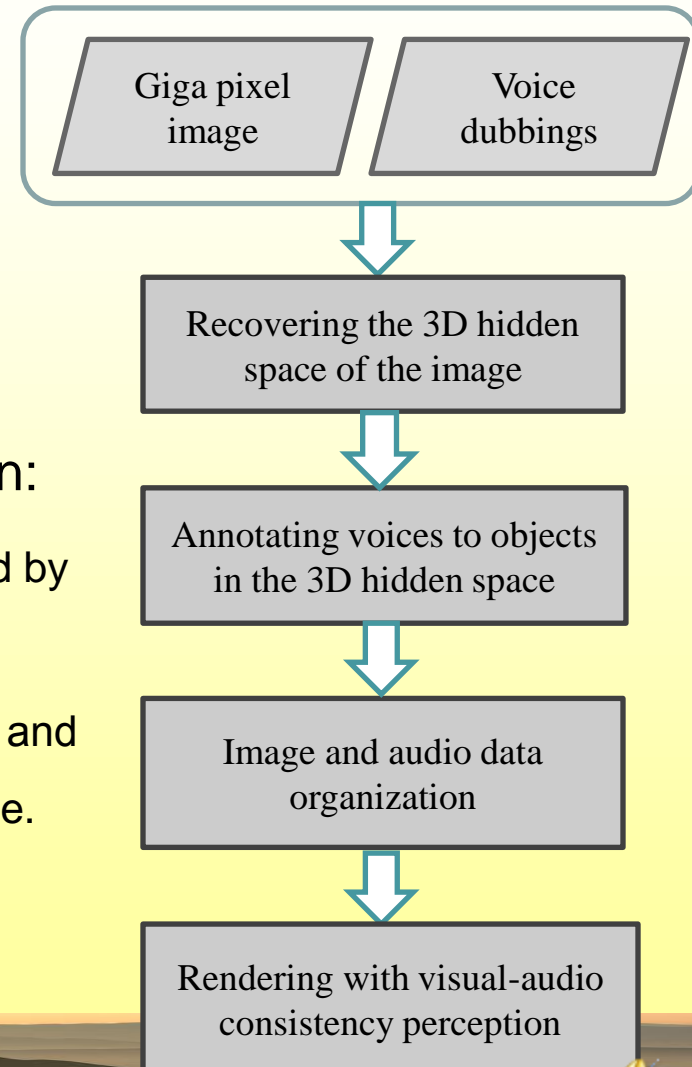# Annotating Chinese Scrolls for Exhibition

- Enhancing a painting with audio annotations:

  - enhance a giga pixel image of the painting with rich audio annotations (the voices of the objects in the painting) ;

- Providing visual-audio consistency perception:

  - put the annotations in the 3D hidden space depicted by the painting;

  - render stereo audio fields by the viewer's distances and orientations to the annotated objects in the 3D space.

```
┌─────────────────────────────┐
│  Giga pixel      Voice      │
│  image           dubbings   │
└─────────────────────────────┘
              ↓
┌─────────────────────────────┐
│  Recovering the 3D hidden   │
│  space of the image         │
└─────────────────────────────┘
              ↓
┌─────────────────────────────┐
│  Annotating voices to objects│
│  in the 3D hidden space     │
└─────────────────────────────┘
              ↓
┌─────────────────────────────┐
│  Image and audio data       │
│  organization               │
└─────────────────────────────┘
              ↓
┌─────────────────────────────┐
│  Rendering with visual-audio│
│  consistency perception     │
└─────────────────────────────┘
```

Annotating Chinese Scrolls for Exhibition

- Demo

Thanks!

Huang Tiejun
tjhuang@pku.edu.cn
http://www.ulib.org.cn
http://idm.pku.edu.cn