Overview of NTCIR-11

Hideo Joho

University of Tsukuba

Kazuaki Kishida

Keio University

Outline
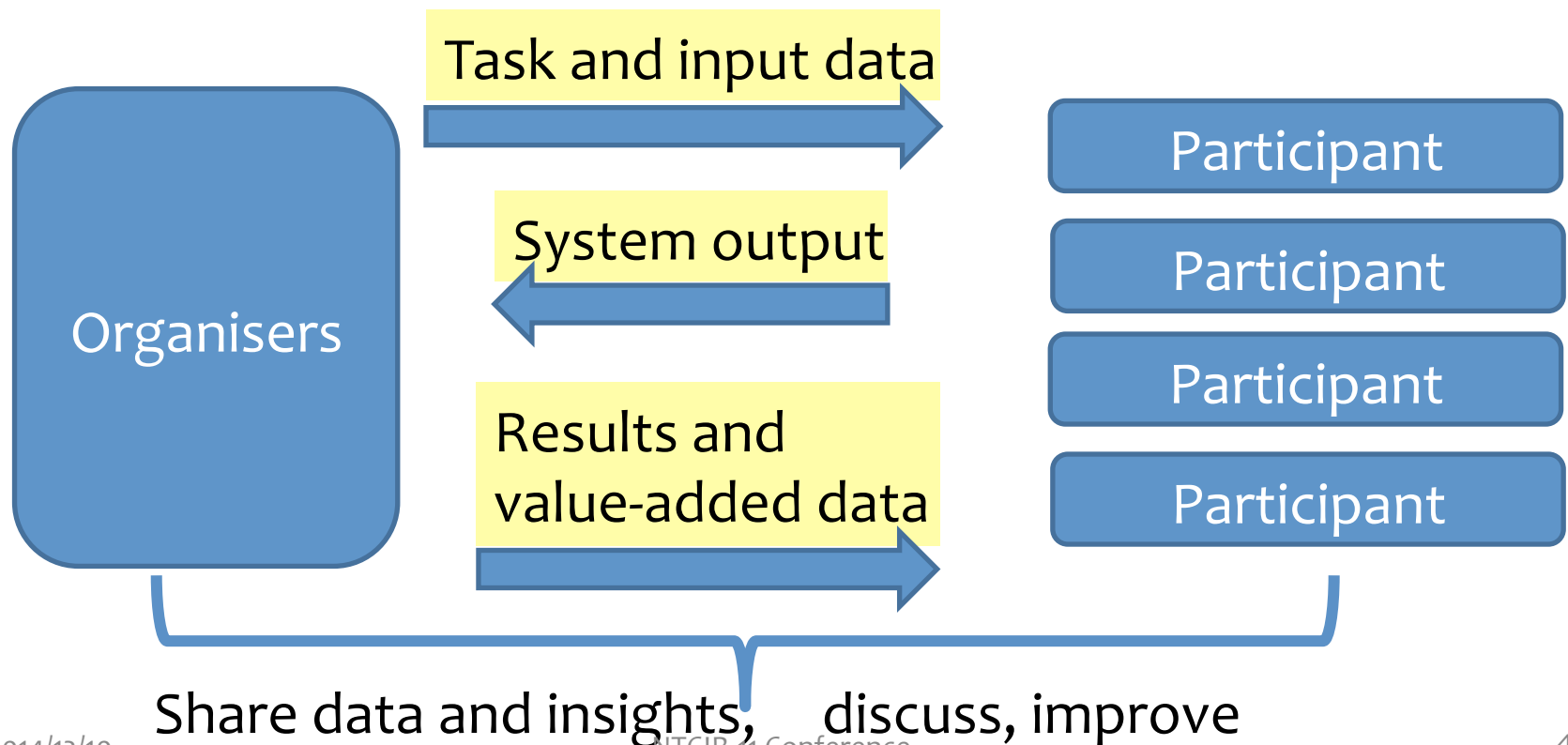
- Basic Concepts of Test Collection Based Evaluation

- Overview of NTCIR-11

- NTCIR-11Evaluation Tasks

- Wrap up

Quick Tutorial

Basic Concepts of Test Collection Based Evaluation

Evaluation Forums

- Research teams gather up to solve shared problems; submit system output before deadline

- Systems evaluated and compared across teams

Task and input data

Organisers

System output

Participant

Participant

Participant

Participant

Results and value-added data

Share data and insights,   discuss, improve

Why Evaluation Forums?

- Compete and collaborate, accelerate research

- Build large-scale test collections through collective efforts e.g. *pooling*

- Foster interdisciplinary research towards grand challenges, build new research communities

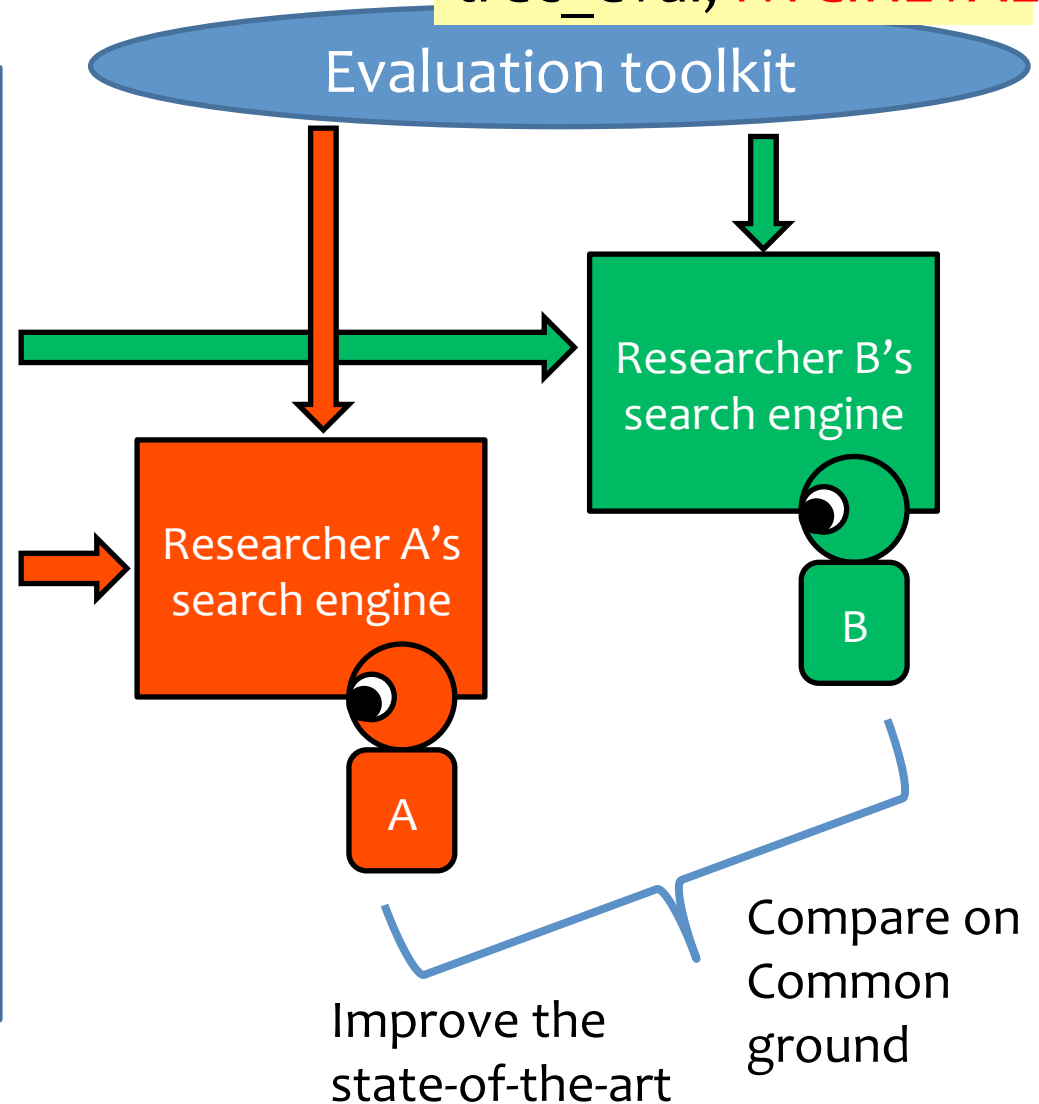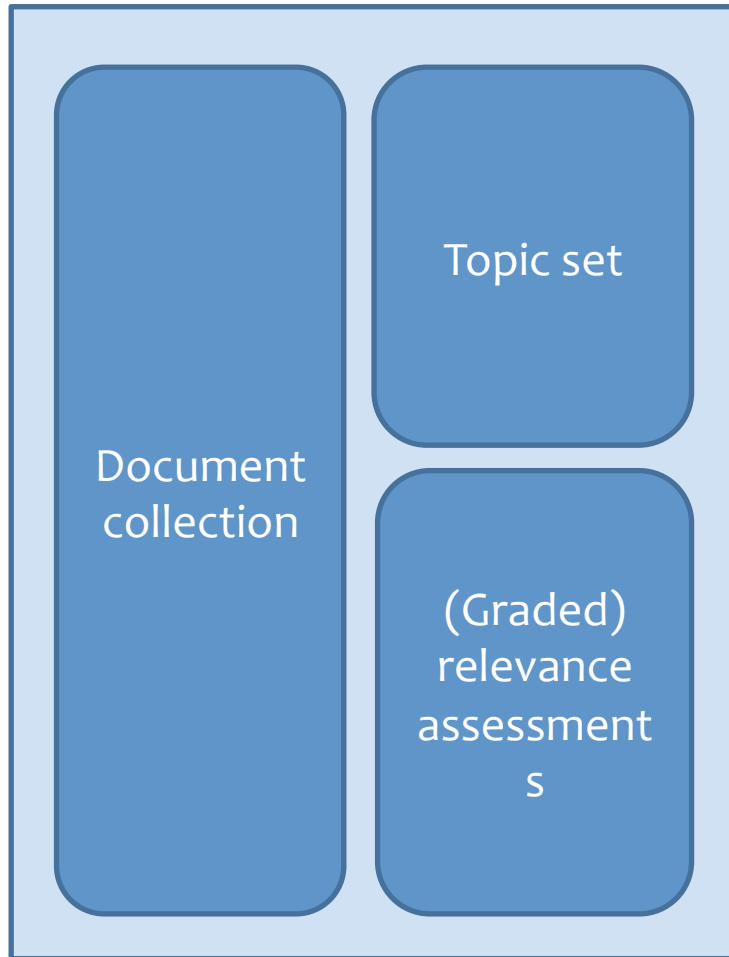NII Testbeds and Community for Information access Research

Information Retrieval Evaluation Forums

- TREC (Text Retrieval Conference) 1992-
- NTCIR 1998/9- [sesquiannual]
- CLEF (Cross-Language Evaluation Forum) 2000-
- INEX (Initiative for the evaluation of XML retrieval) 2002-
- TRECVID 2003-
- FIRE (Forum for IR Evaluation) 2008-
- MediaEval (Benchmarking Initiative for Multimedia Evaluation) 2010-

[not exhaustive]

Jargon: Test Collection
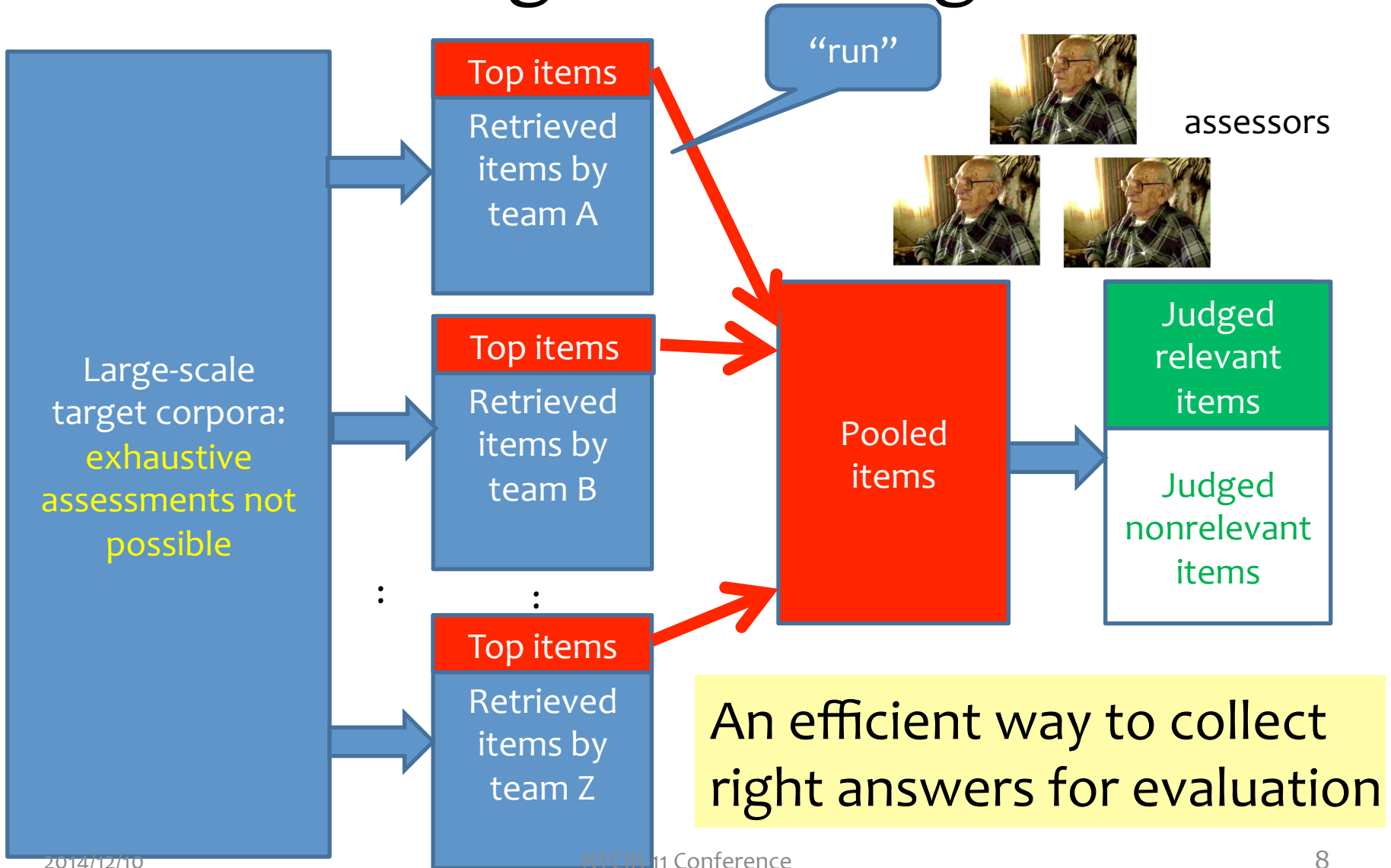

Test collection

trec_eval, NTCIREVAL
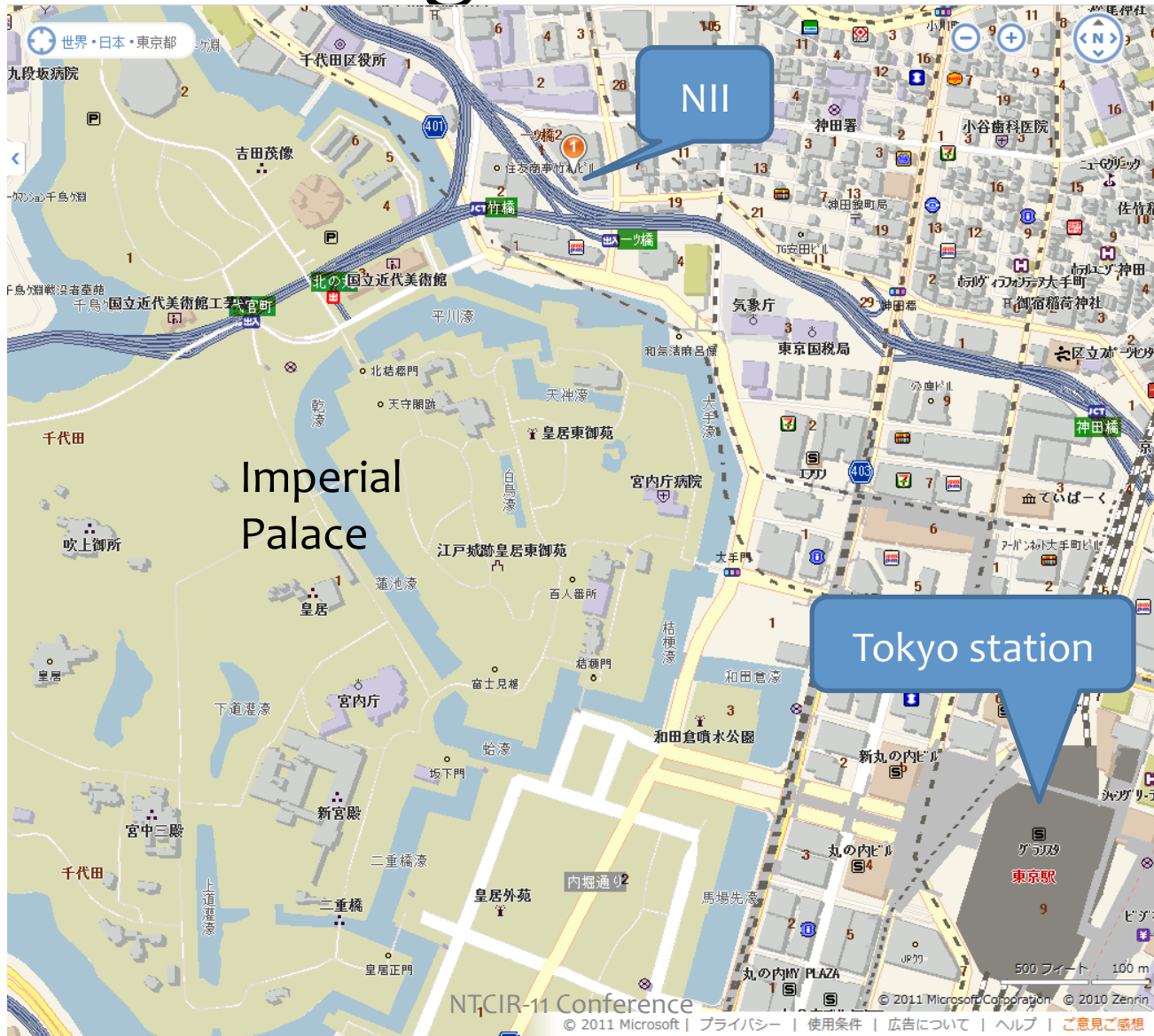
Evaluation toolkit

Document collection

Topic set

(Graded) relevance assessments

Researcher A's search engine

A

Researcher B's search engine

B

Improve the state-of-the-art

Compare on Common ground

Jargon: Pooling

Large-scale target corpora: exhaustive assessments not possible

Top items
Retrieved items by team A

Top items
Retrieved items by team B

Top items
Retrieved items by team Z

"run"

assessors

Pooled items

Judged relevant items

Judged nonrelevant items

An efficient way to collect right answers for evaluation

Jargon: MAP

Jargon: MAP

Query:
**(I want to eat) sushi**

寿司 - Wikipedia
語源説・種類・地方の寿司・歴史
寿司（すし、鮨、鮓、寿斗、寿し、壽司）と呼ばれる食品は、酢飯 と主に 魚介類 を組み合わせた日本料理 である。 大別すると、生鮮魚介を用いた「早鮨（早ずし）」と、魚介類に米を加えて乳酸発酵させた「なれ鮨（なれずし）」に区分される。 そのなかでも代表的な寿司は前者の 握り寿司（江戸前 寿司）であり、すでに" sushi "で通じるほど世界中に …
ja.wikipedia.org/wiki/寿司・スパムとして報告
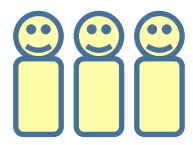
元気寿司グループ | 回転寿司のパイオニア
11/09/14 千両・廻鮮日本海【秋期間限定メニュー】販売中！！ 11/09/01 すしおんど【秋期間限定メニュー】販売中！！ 9月1日（木)より 11/09/01 元気寿司【秋期間限定メニュー】販売中！！ 9月1日（木 …
www.genkisushi.co.jp/index.php・スパムとして報告
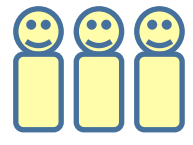
かっぱ寿司へようこそ!!
かっぱ寿司で使用している国産の魚介類産地について かっぱ寿司で使用している魚介類の産地につきましては、「原産地情報」ページで日々最新情報をご提供しております。 厳選素材の美味しさをお楽しみください「秋のデカ旨」開催中！
www.kappa-create.co.jp・スパムとして報告
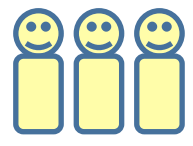
寿司(鮨)ランキング [食べログ]
全国にある寿司(鮨)のお店32,640件を一般ユーザーの口コミをもとに集計した様々なランキングから探すことができます。美味しい寿司(鮨)のお店や、人気の寿司(鮨)のお店が簡単に見つかります！
r.tabelog.com/sushi・スパムとして報告
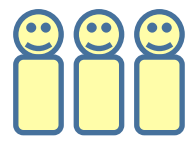
東京の寿司 マグロが自慢【すしざんまい】tsukiji sushi-zanmai
お寿司といえば『すしざんまい』！！江戸前寿司の本場、東京で美味しい寿司が食べたいそんなあなた、『すしざんまい』へお越しください！tsukiji sushi-zanmai … 《 What's NEW 》 特選おせち【三段重】予約申込受付中！ 11/13（日 …
www.kiyomura.co.jp・スパムとして報告

寿司 レシピ 689品 [クックパッド] 簡単おいしいみんなのレシピが109万品
人気の寿司レシピ（作り方）が689品集まった寿司コミュニティ。毎日更新の寿司レシピランキング みんなが料理した写真付きレポートなどが充実している。 … 寿司 推薦レシピ：689 品 つくれ…件 手作り寿司のレシピ。お寿司のアイディア …
…tegory/165・スパムとして報告

Precision@2=1/2

Precision@3=2/3

Precision@4=3/4

Precision@5=4/5

$$AP = \frac{\sum_r \text{Isrelevant}(r) * \text{Prec@}r}{\#\text{all relevant docs}}$$
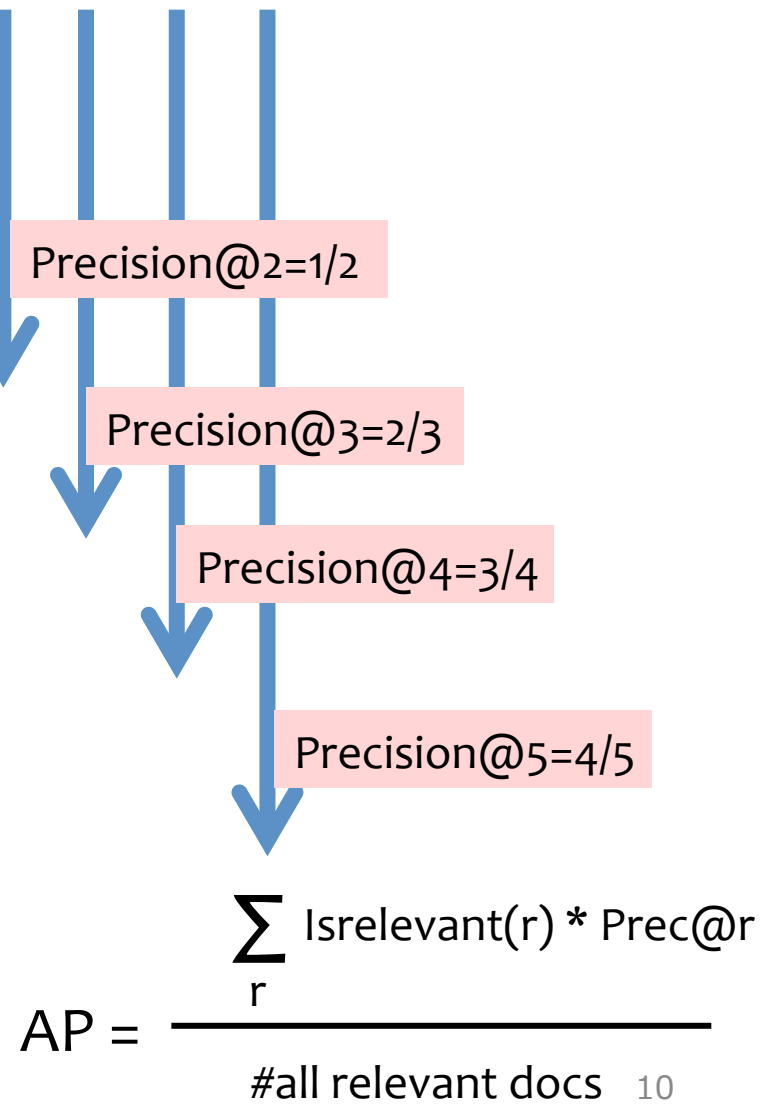
User's stopping probability: uniform

Jargon: nDCG

normalised
Discounted
Cumulative
Gain

System's output — **D**iscounted **G**ain

System's output	Discounted Gain
Nonrelevant	
Relevant	$2/\log(2+1)$
Relevant	
Partially relevant	$1/\log(4+1)$

Ideal output — **D**iscounted **G**ain

Ideal output	Discounted Gain
Highly relevant	3
Relevant	$2/\log(2+1)$
Relevant	$2/\log(3+1)$
Partially relevant	$1/\log(4+1)$

$$\text{nDCG} = \frac{\sum \text{System's discounted gains}}{\sum \text{Ideal discounted gains}}$$

Unlike AP,
nDCG can utilise
graded relevance assessments
(widely used in Web search
evaluation)

Recommended Books

- Sanderson, M. (2010) Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, 4(4), pp.247-375. http://www.seg.rmit.edu.au/mark/fntir/

- 酒井哲也 (2015) 情報アクセス評価方法論～検索エンジンの進歩のために～. コロナ社. 2015年夏刊行予定！

Introduction

Overview of NTCIR-11

NTCIR-11 Tasks

[Core Tasks]

- IMine: Search Intent and Task Mining

- Math-2: Mathematical Information Access

- MedNLP-2: Medical Natural Language Processing

- MobileClick: Mobile Information Access

- RITE-VAL: Recognizing Inference in Text and Validation

- SpokenQuery&Doc: Spoken Query and Spoken Document Retrieval

[Pilot Tasks]

- QALab: QA Labs for Entrance Exam

- RecipeSearch: Cooking Recipe Search

- Temporalia: Temporal Information Access

Program Committees

- **Hsin-Hsi Chen** | National Taiwan University, Taiwan

- **Charles Clarke** | University of Waterloo, Canada

- **Kalervo Järvelin** | University of Tampere, Finland

- **Gareth Jones** | Dublin City University, Ireland

- **Gary Geunbae Lee** | POSTECH, South Korea

- **Maarten de Rijke** | University of Amsterdam, The Netherlands

- **Stephen Robertson** | Microsoft Research Cambridge, UK

- **Ian Soboroff** | NIST, US

- **Hideo Joho** Co-chair | University of Tsukuba, Japan

- **Kazuaki Kishida** Co-chair | Keio University, Japan

Historical Development

Year	1999	2001	2002	2004	2005	2007	2008	2010	2011	2013	2014
Task/NTCIR round	1	2	3	4	5	6	7	8	9	10	11
Automatic Term Recognition and Role Analysis (TMREC)	9										
Ad hoc/Crosslingual IR(1) -> Chinese/English/Japanese IR(2) -> CLIR(3-6)	28	30	20	26	25	22					
Text Summarization Challenge (TSC)		9	8	9							
Web Retrieval (WEB)			7	11	7						
Question Answering Challenge (QAC)			16	18	7	8					
Patent Retrieval [and Classification] (PATENT)			10	10	13	12					
Multimodal Summarization for Trend Information (MUST)					13	15	13				
Crosslingual Question Answering (CLQA)(5,6) -> Advanced Crosslingual Information Access (ACLIA)(7,8)					14	12	19	14			
Opinion(6) -> Multilingual Opinion Analysis (MOAT)(7,8)						12	21	16			
Patent Mining (PAT-MN)							12	11			
Community Question Answering (CQA)								4			
Geotemporal IR (GeoTime)								13	12		
Interactive Visual Exploration (Vis-Ex)									4		
Patent Translation (PAT-MT)(7,8) -> Patent Machine Translation (PatentMT)(9,10)							15	8	21	21	
Crosslingual Link Discovery (Crosslink)									11	10	
INTENT(9,10) -> **Search Intent and Task Mining (IMine)**									16	11	**12**
One Click Access (1CLICK)(9,10) -> **Mobile Information Access**									4	8	**4**
Recognizing Inference in Text (RITE)(9,10) -> **Recognizing Inference in Text and Validation (RITE-VAL)**									24	28	**23**
IR for Spoken Documents (SpokenDoc) (9,10) -> **Spoken Query and Spoken Document Retrieval**									10	12	**11**
Mathematical Information Access (Math)										6	**8**
Medical Natural Language Processing (MedNLP)										12	**12**
QA Lab for Entrance Exam (QALab)											**11**
Temporal Information Access (Temporalia)											**8**
Cooking Recipe Search (RecipeSearch)											**4**
	37	39	61	74	79	81	80	66	102	108	**93**

Why Declined?

- New (young) tasks tend to have smaller user-base than established (mature) tasks

- Advanced task design might have raised the bar little too high

- Available resources have been limited in many organisations

- Yet, still the 3rd largest participation

Scale of the community

- 9 Evaluation Tasks

- 47 Task Organisers

- 100+ Papers in NTCIR-11 Proceedings

- 323 Authors

- 20 countries/regions

Languages in NTCIR-11

	IMine	Math-2	Spoken Query& Doc	QALab	Recipe Search	MedNLP-2	Mobile Click	RITE-VAL	Temporalia
Chinese		■	■	■	■	■	■		■
English			■			■			
Japanese		■							■

Geographic Distribution

Table 1: Number of participating teams by country/region

Country/Region	# of Teams
Japan	42*
China	12*
Taiwan	9
Germany	4
USA	4
Korea	3*
Ireland	2
Australia	1*
Austria	1
Bangladesh	1*
Canada	1
Czech Republic	1
France	1
India	1
Norway	1
Sweden	1*
United Kingdom	1
Vietnam	1*
TOTAL	87 (81)

Countries with * had joint international teams.

Sneak Preview

NTCIR-11 Evaluation Tasks

IMine

- Organisers: Tsinghua Univ., MSRA, Kyoto Univ., Edinburgh Univ.

- Many search queries are short, ambiguous, under-specified

- Disambiguation is difficult when underlying search intents can be diverse

- An alternative solution is to diversify search results based on potential search intents

IMine

- Subtasks
  - Subtopic Mining: Mining of candidate intents for a given query [CJE]
  - Document Ranking: Generation of diversified ranked list of documents [CE]
  - Task Mining: Automatic finding of subtasks of a parent task (e.g., moving) [J]

Math-2

- Organisers: NII, Jacobs Univ. Bremen, Glasgow Univ., Tech Univ. Berlin

- Mathematical formula is a universal language across many scientific disciplines

- Yet, little work has been done to allow researchers to access documents via mathematical representations

- Math-2 is the 2nd running of mathematical information access challenges

Math-2

- Subtasks
  - Ad-hoc Retrieval Task: Generation of ranked list of formula containing texts that match a query [E]

  - Wikipedia Open Task: Researchers can pursue their own agenda using math-related articles in Wikipedia [E]

MedNLP-2

- Organisers: Kyoto Univ., Tokyo Univ., Shizuoka Univ., Fuji Xerox

- Digital medical records have increased and started to replace paper-based documents

- A great opportunity to help both medical practitioners and patients by developing IA systems for medical information

- MedNLP-2 is the 2nd running of NLP challenges in Medical domain

MedNLP-2

- Subtasks
  - Extraction Task: Extract disease names, date/time from medical reports [J]
  - ICD Coding Task: Assignment of standardised classification codes to disease names [J]
  - Free Task: Researchers can pursue their own agenda using the collection [J]

MobileClick

- Organisers: Kyoto Univ., Northeastern Univ., Waseda Univ., KDDI Corp.

- Although advance of mobile devices is remarkable, there is a still large difference from conventional desktop PCs

- A key technology to support mobile users is search results summarisation

MobileClick

- Subtasks
  - iUnit retrieval task: Generation of ranked list of iUnits for a given query [EJ]
  - iUnit summarisation task: Two-layers of textual outputs as a summary [EJ]

RITE-VAL

- Organisers: Yamanashi Univ., NII, Kyoto Univ., National Taiwan Ocean Univ., Academia Sinica, NEC Corp., CMU

- Some IA system such as QA can benefit from deeper text analysis

- Text entailment is a process of determining the relationship between two pairs of texts

RITE-VAL

- Subtasks
  - Fact validation task: Determine if a text can be inferred from another text [CJE]
  - System validation task: Detailed analysis of effects of linguistic features on system behaviour [CJ]

SpokenQuery&Doc

- Organisers: Toyohashi Univ., Yamanashi Univ., Ryukoku Univ., DCU

- Much of IA technologies is based on written documents

- However, archiving of spoken documents has increased, audio inputs have become common

SpokenQuery&Doc

- Subtasks
  - Spoken-query driven spoken content retrieval: Retrieval of spoken documents for a given spoken query [J]

  - Spoken-query driven spoken term detection: Detection of a given spoken term in spoken documents [J]

  - Spoken term detection based spoken content retrieval: Retrieval of spoken documents using terms extracted from spoken queries [J]

QALab

- Organisers: Yokohama National Univ., NII, Shizuoka Univ., CMU

- When the complexity of IA system increases, multiple components are needed

- Real-world QA systems are one such case

- Optimal integration and evaluation of multiple components is not trivial

QA Lab

- Modules
  - Question Analysis
  - Document Retrieval
  - Candidate Answer Extraction
  - Answer Generation
- Combinations

RecipeSearch

- Organisers: Gunma Univ., Microsoft Research, Yummly, Kagawa Nutrition Univ.

- Cooking is a domain where Internet has drastically changed the way we seek information

- Cooking is such a creative activity that there can be many recipes to cook the same dish

RecipeSearch

- Subtasks
  - Ad-hoc search task: Retrieval of recipes for a given query [EJ]
  - Recipe pairing task: Retrieval of side dish recipes for a given main meal [EJ]

Temporalia

- Organisers: Tsukuba Univ., Kyoto Univ., Yahoo Lab

- Many search queries have temporal intents/ implications across a range of classes (e.g., past, recent, future)

- Yes, no test collection is available to evaluate temporal IR systems in a structured way

Temporalia

- Subtasks

  - Temporal Query Intent Classification: Classification of queries into temporal classes [E]

  - Temporal Information Retrieval: Generation of ranked list of documents for each of temporal subtopics [E]

Highlights

Highlights of NTCIR-11

- A good combination of refined tasks and brand new tasks

- Broad technical challenges: Retrieval, classification, summarisation, term detection, textual inference, QA, NER

- Rich domains: Web pages, News, Medical, Math, Exams, Lectures, Recipes

- Excellent task organising teams!

Wrap-up

Wrap-up

- 11th running of NTCIR

- 9 exciting tasks: Refined & brand new

- Contributions from broad communities

- Keynote, task overviews, participants' oral presentations & poster presentations

- Online proceedings are available

Many thanks to

Assessors

NTCIR Office

Task Organisers

Task Participants

Conference Sponsors

Conference Attendees

and

National Institute of Informatics

Enjoy the conference!