Overview of the NTCIR-10 1CLICK-2 Task

Makoto P. Kato

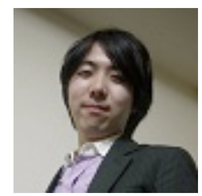
Matthew Ekstrand-Abueg

Virgil Pavlu

Tetsuya Sakai

Takehiro Yamamoto

Mayu Iwata



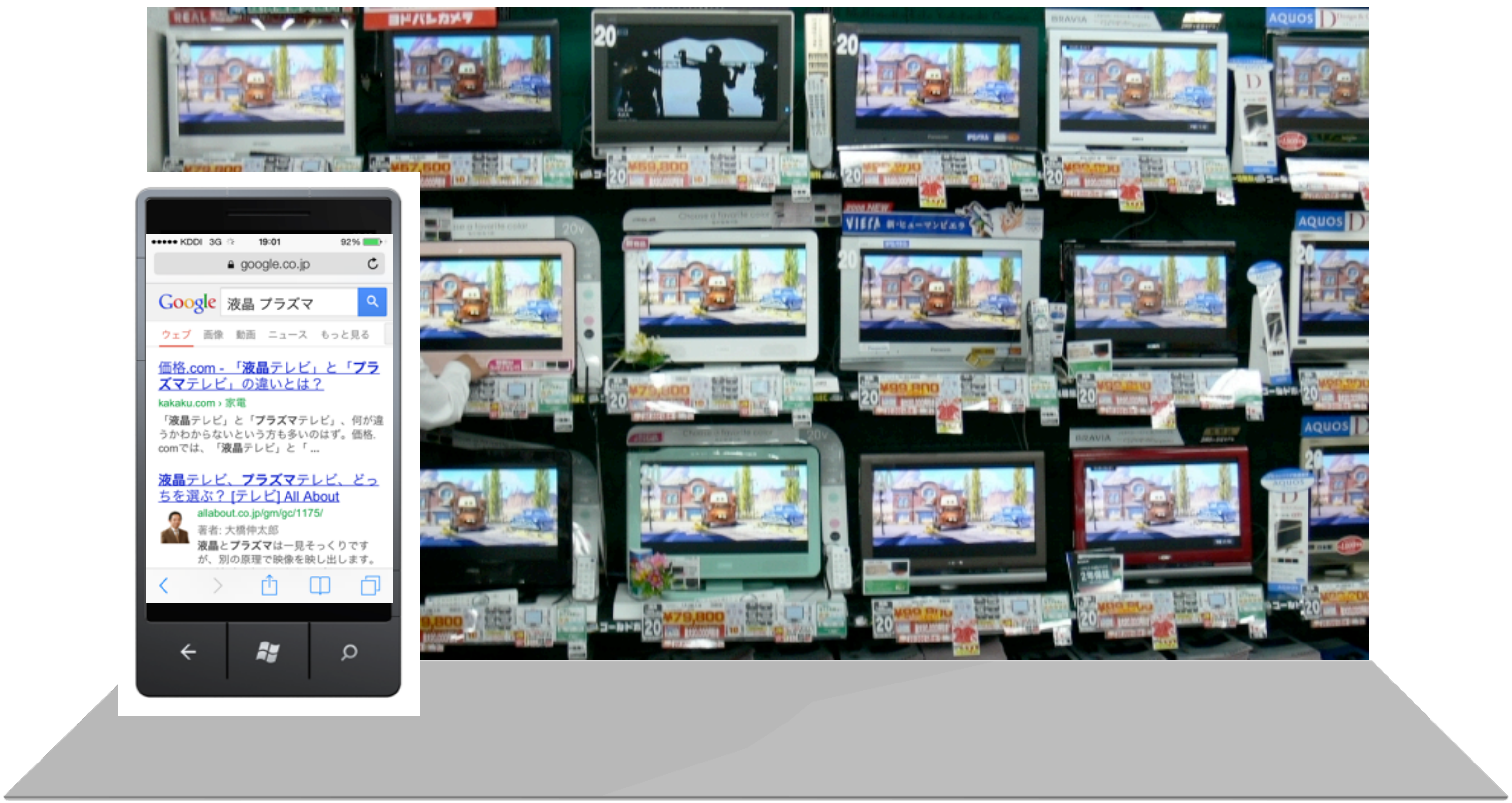
Talk Outline

1. What is 1CLICK Task?
2. The 1CLICK-2 Task
3. Results
4. Summary and Future work

What is the 1CLICK Task?

Suppose that …

- Finding answers for a question
  “what’s the difference between PDP and LCD?”

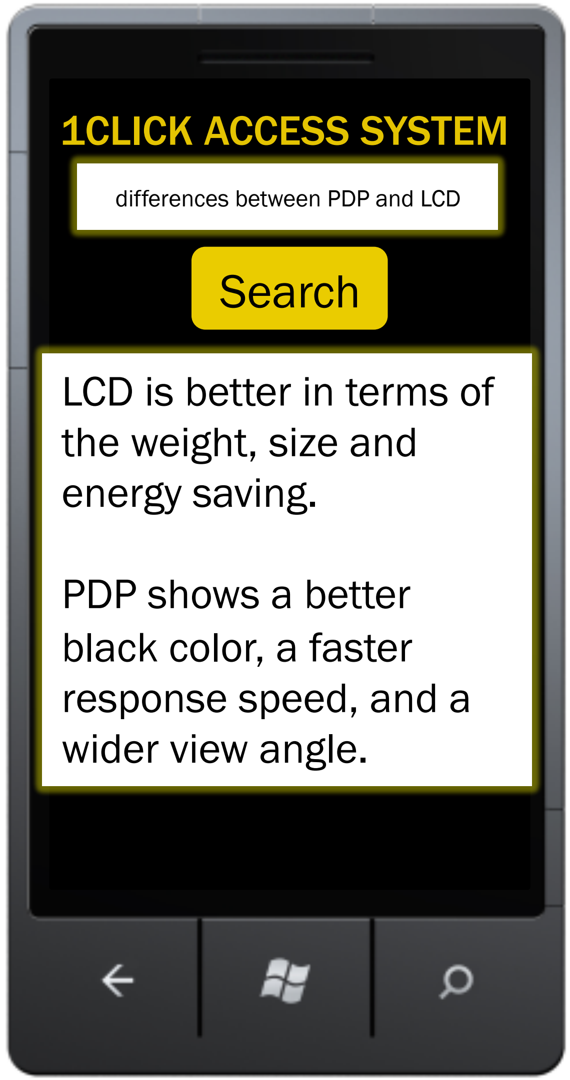


In the "ten-blue-link" paradigm

- Enter query
- Click SEARCH button
- Scan ranked list of URLs
- Click URL
- Read URL contents
- Get all desired information

More than one clicks needed before being satisfied

This is "One Click Access"



One Click Access

=

Immediate

+

Direct

Information Access

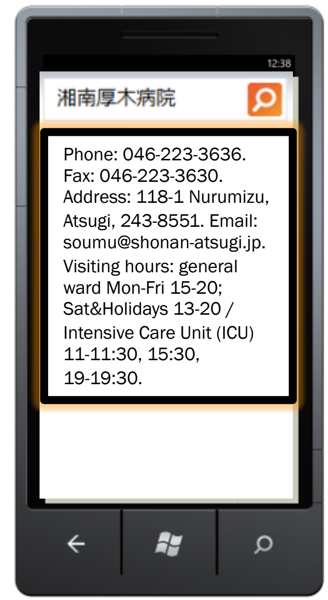
One Click Access

Enter query

Click SEARCH button

Get all desired information

The system outputs *X-string*



Task:
Given a search query, return
a single textual output (*X-string*)

Go beyond the "ten-blue-link" paradigm, and tackle *information* retrieval rather than document retrieval

Evaluation of 1CLICK Systems

- Manual/automatic matching between the **X-string** and **nuggets**

<u>Phone: 046-223-3636</u>. Fax: 046-223-3630.
Address: 118-1 Nurumizu, Atsugi, 243-8551.
Email: soumu@shonan-atsugi.jp

X-string

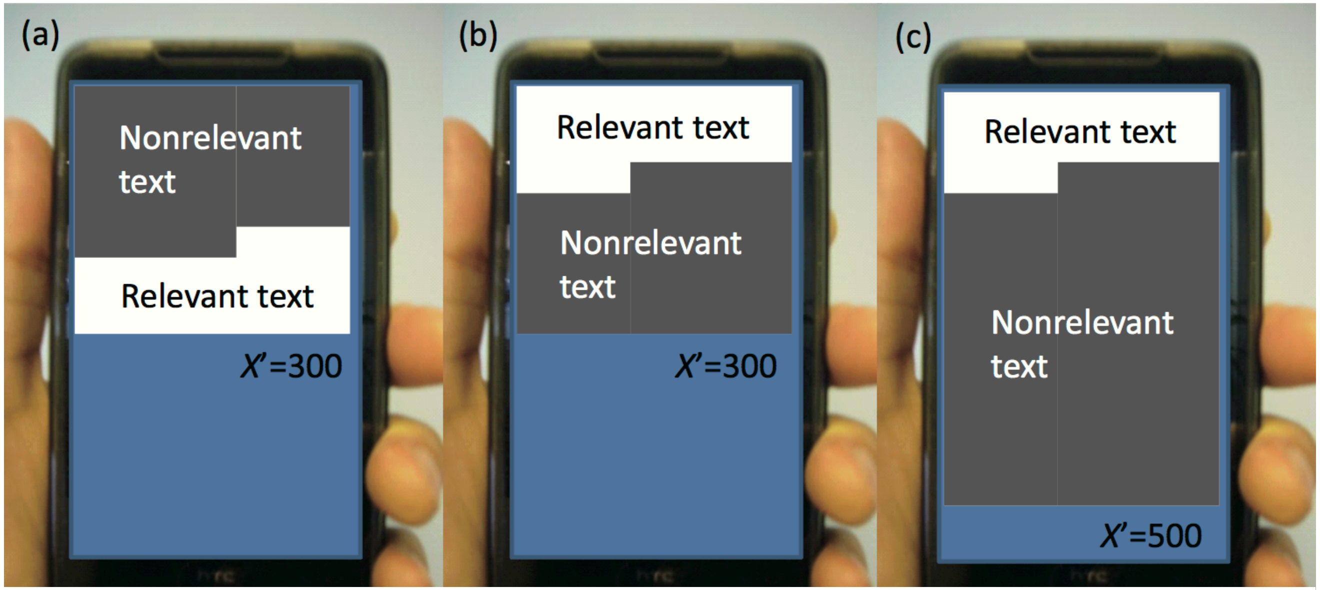
- <u>Phone number: 046-223-3636</u>
- Fax number: 046-223-3630
- Address: 118-1 Nurumizu, Atsugi

Nuggets

a sentence relevant to the information need for a query

- Systems are required to present <u>more important information earlier</u>

Evaluation Metrics for 1CLICK



(a) (b) (c)

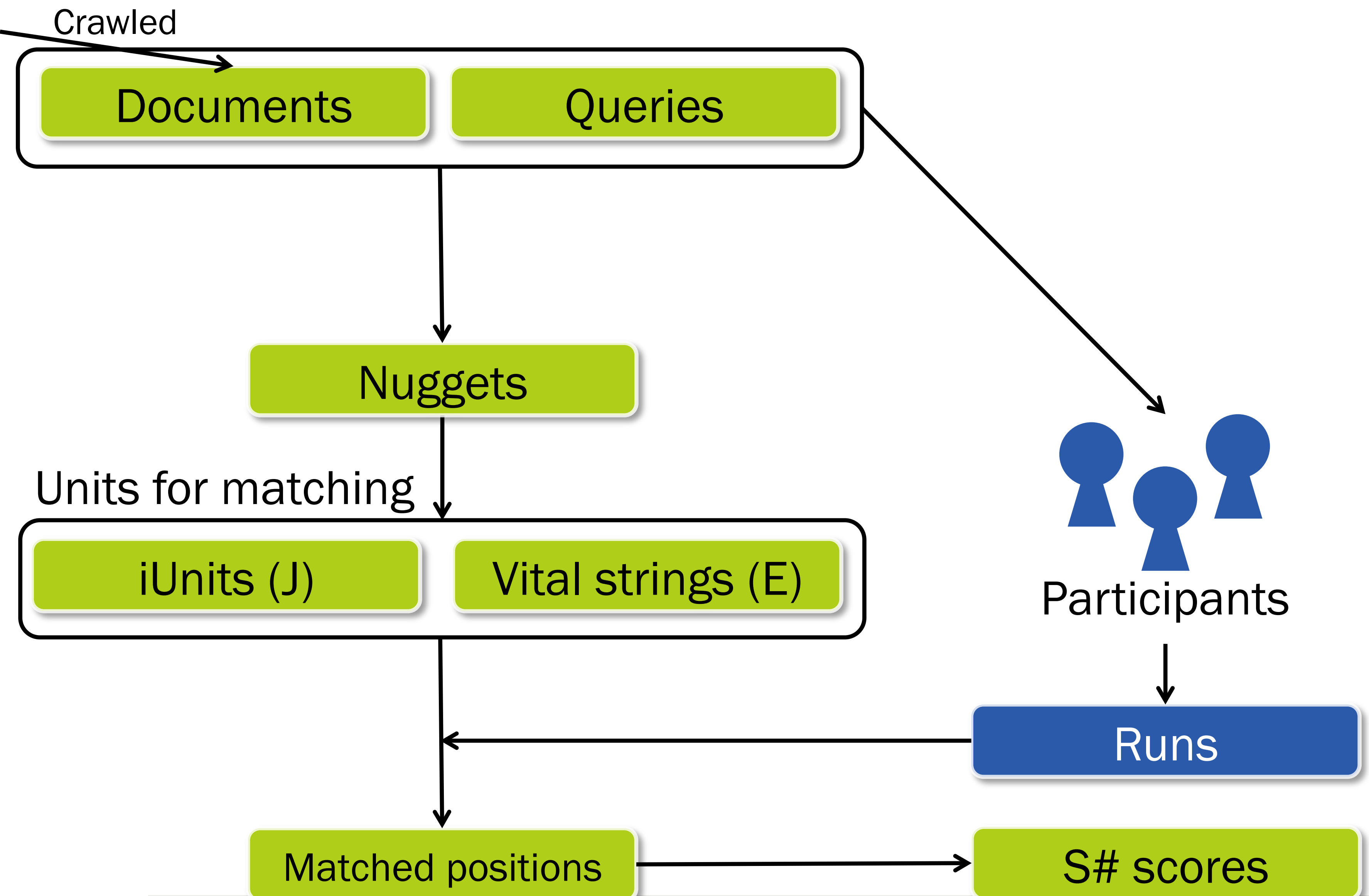
- Unlike nugget precision/recall, S-measure (position-aware weighted recall) says (a)<(b). T-measure (a kind of precision) says (b)>(c). S# (official evaluation metric) combines S and T

1CLICK Challenges

- For participants
  - Multi-document summarization for a given query
  - Precise estimation of the nugget importance
    - Not binary but graded importance
  - Readability of X-strings
- For organizers
  - Efficient nugget construction
  - Flexible, feasible, and consistent nugget matching
  - Appropriate evaluation metrics

The 1CLICK-2 Task

1CLICK-2 Task Structure



Main tasks (English + Japanese)

- Given a search query,
  return a X-string (a single textual output)
- Options
  - Device types (the length limit for X-strings)
    - DESKTOP: 1,000/500 characters for English/Japanese
    - MOBILE: 280/140 characters for English/Japanese
  - Source types NEW
    (from which X-strings must be generated)
    - MANDATORY: only distributed documents
    - ORACLE: only distributed documents with an “ORACLE” list
    - OPEN: any resources

Query Classification Subtask (English + Japanese) NEW

- Given a search query, return the query type
  - For *componentized* evaluation

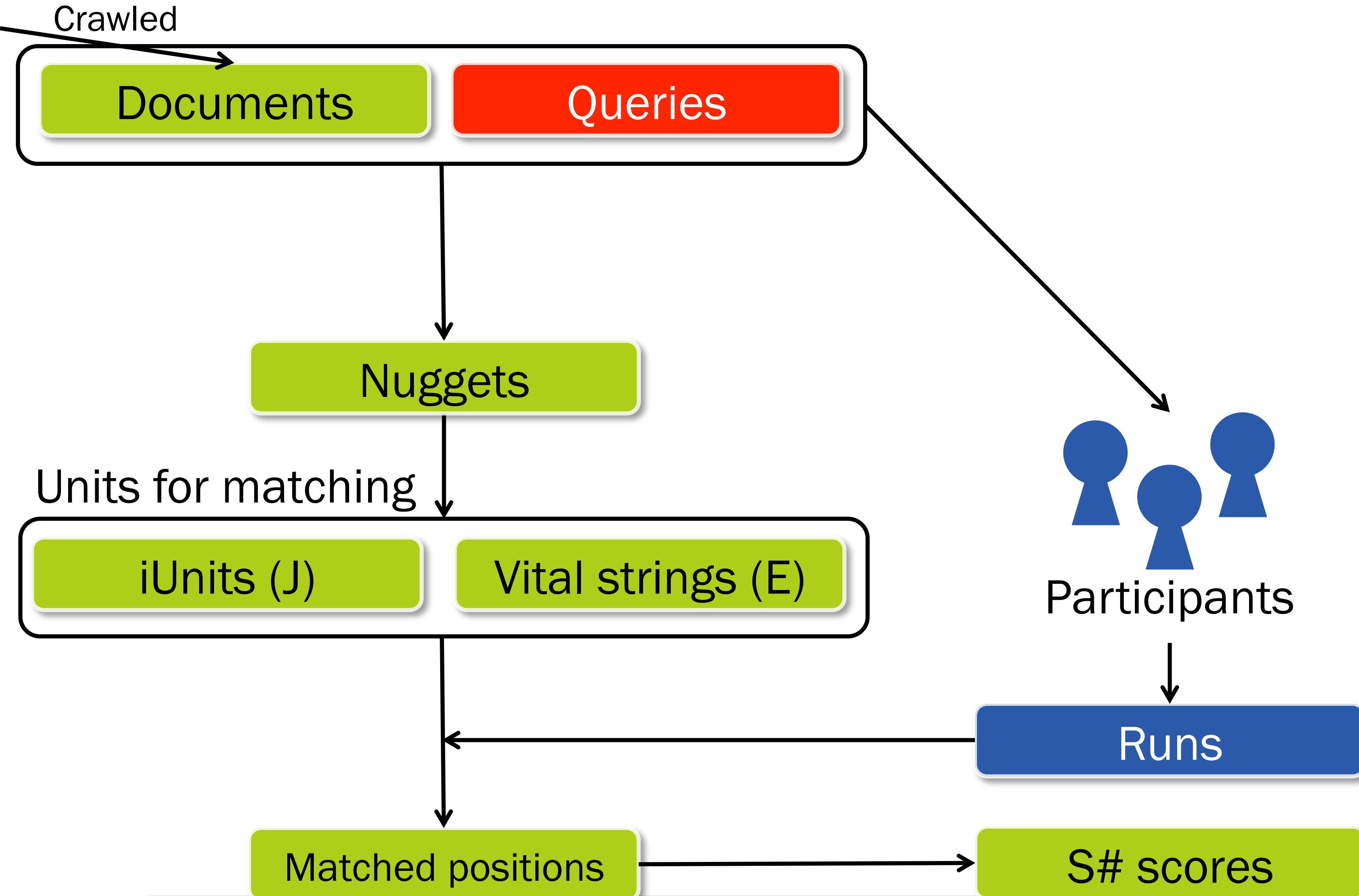
Queries

"michael jackson death"
"sylvester stallone"
"robert kennedy cuba"
"ichiro suzuki"
"atlanta airport"
"kyoto hot springs"
"parkinsons disease"
"why is the sky blue?"

→ Classifier →

Query types
ARTIST
ACTOR
POLITICIAN
ATHLETE
FACILITY
GEO
DEFINITION
QA

1CLICK-2 Task Structure



Queries

NEW

- 8 query types used

(for which the user's information need can be satisfied by the search results page)

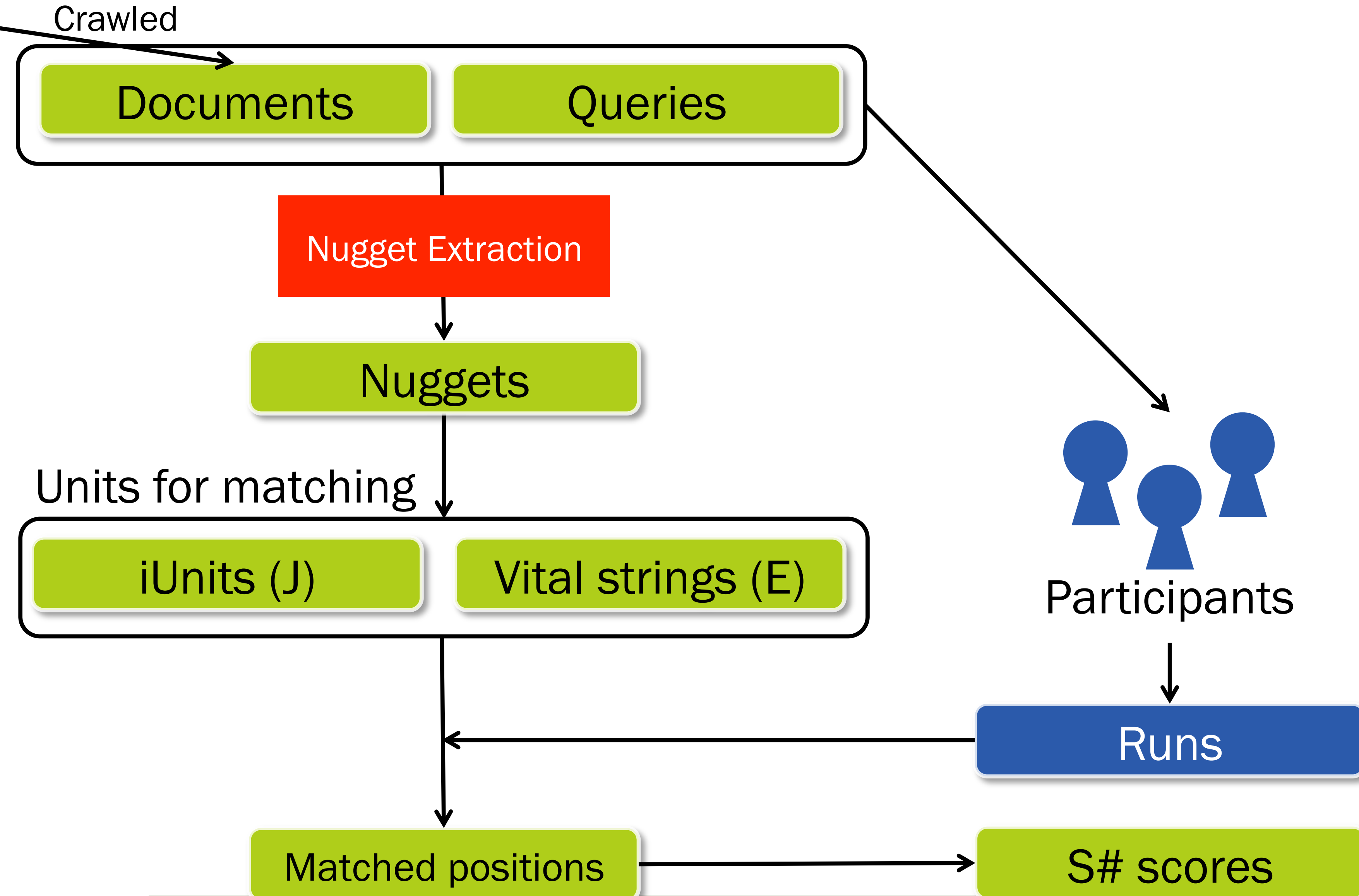
1CLICK-1
CELEBRITY
FACILITY
DEFINITION
QA

→

1CLICK-2	Examples
ARTIST	“michael jackson death”
ACTOR	“sylvester stallone”
POLITICIAN	“robert kennedy cuba”
ATHLETE	“ichiro suzuki”
FACILITY	“atlanta airport”
GEO	“kyoto hot springs”
DEFINITION	“parkinsons disease”
QA	“why is the sky blue?”

Based on [Li et al., SIGIR09]

1CLICK-2 Task Structure



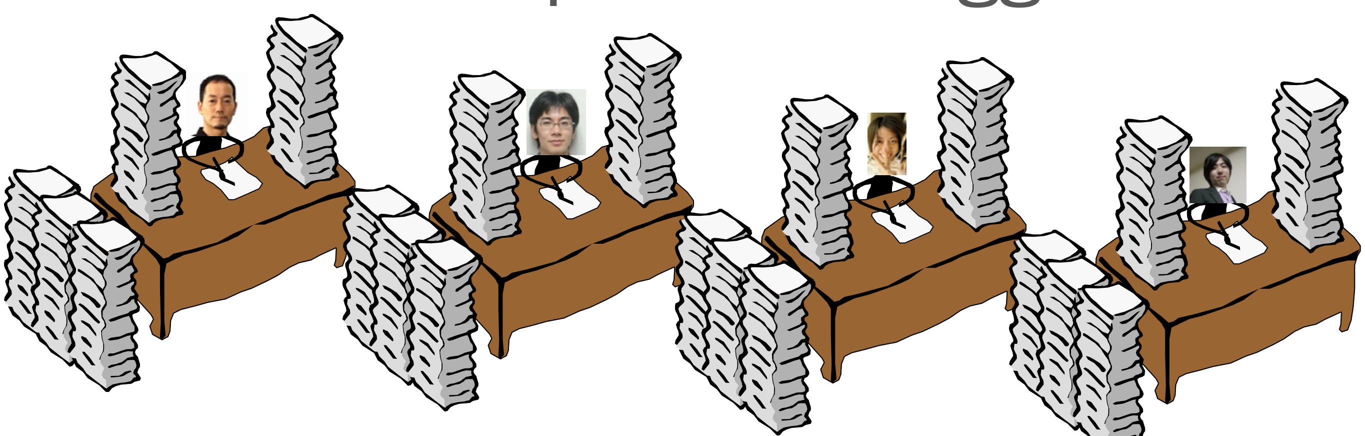
Nugget Extraction

- Nugget: a sentence relevant to a given query
  - e.g. For query "ichiro suzuki",

Nugget

Ichiro is a professional baseball outfielder who is currently with the New York Yankees

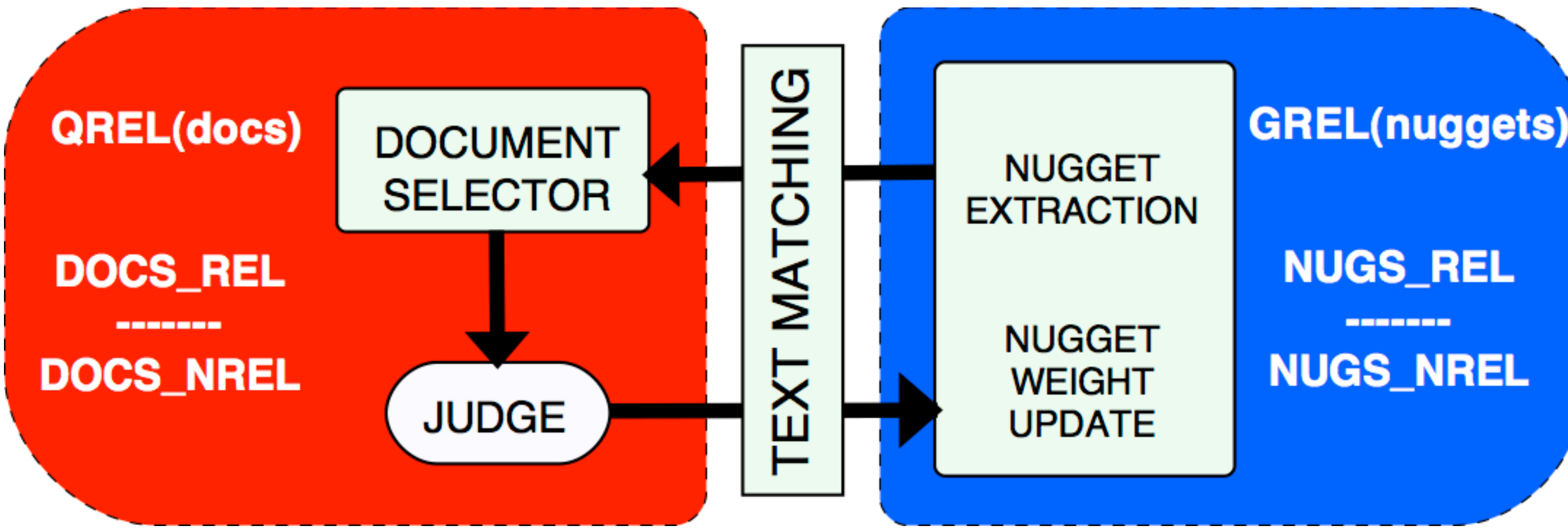
- In Japanese 1CLICK-2, organizers worked very hard to collect all possible nuggets in advance



Very time-consuming process

Semi-automatic Nugget Extraction

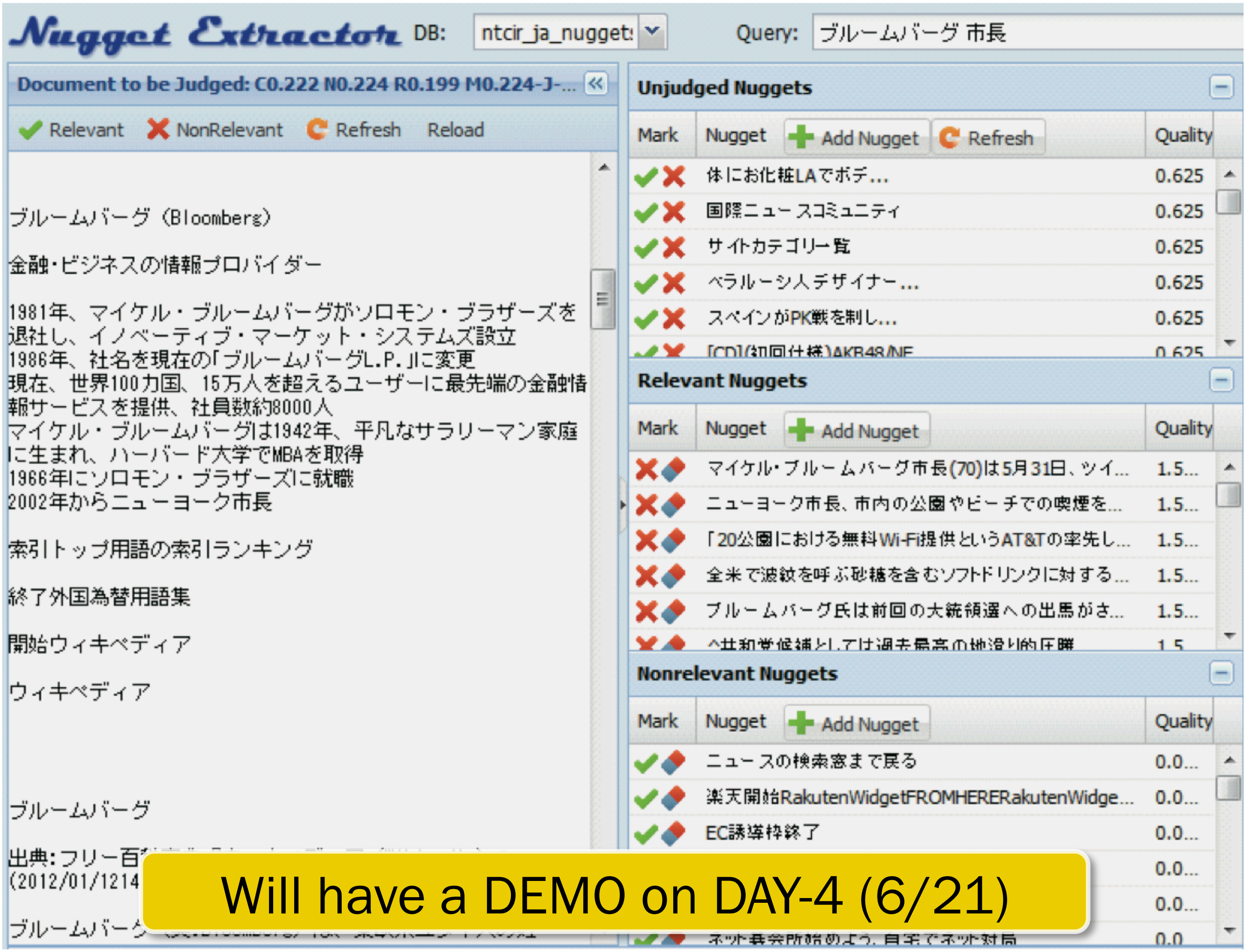

- Was applied to English 1CLICK-2
- Mutual, iterative reinforcement between nuggets and documents



Refer to our SIGIR 2013 paper to see the performance:

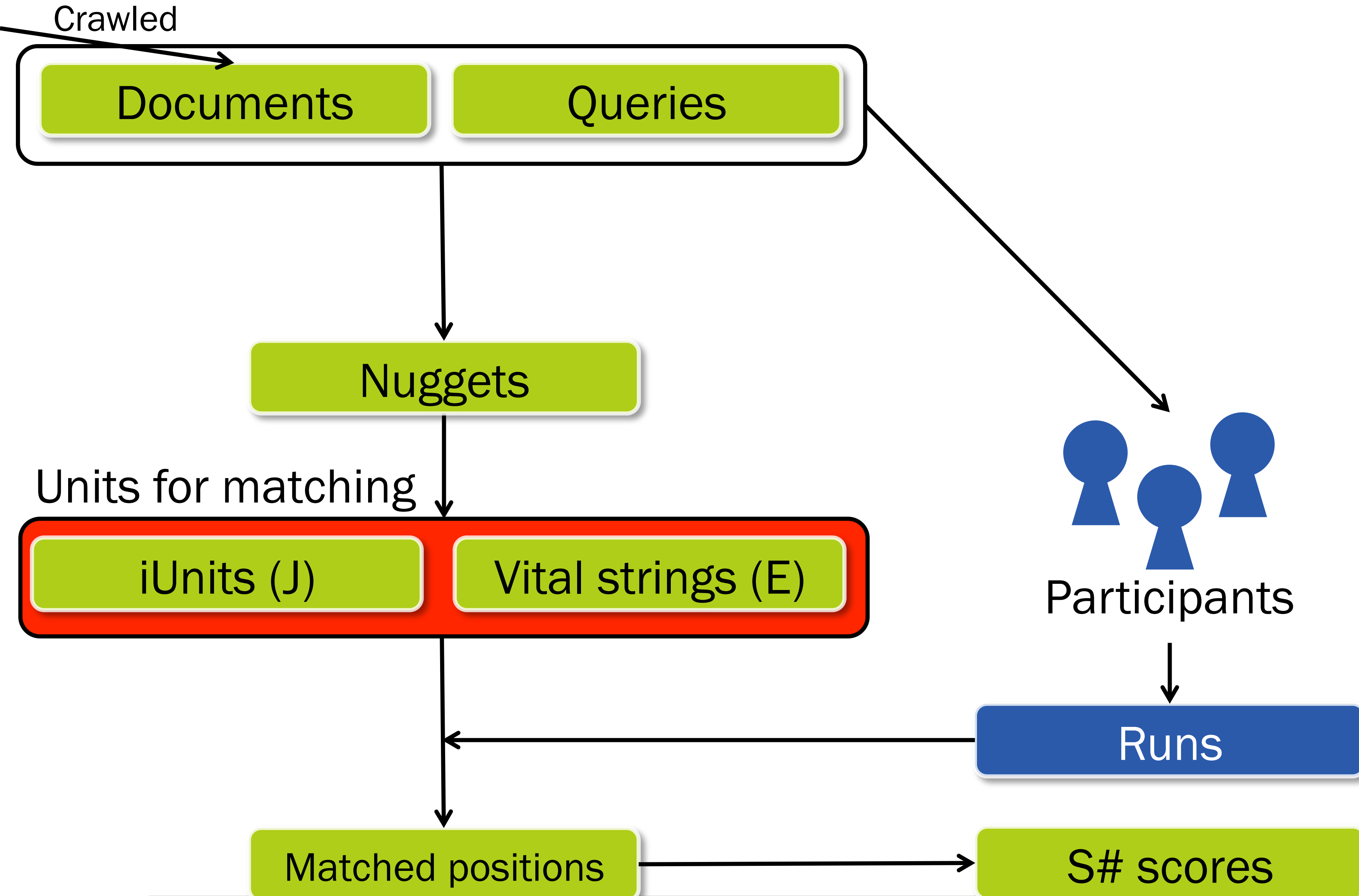
Ekstrand-Abueg, M., Pavlu, V., Kato, M.P., Sakai, T., Yamamoto, T., Iwata, M.: Exploring Semi-Automatic Nugget Extraction for Japanese One Click Access Evaluation, ACM SIGIR 2013, to appear, July 2013.

Nugget Extractor



Will have a DEMO on DAY-4 (6/21)

1CLICK-2 Task Structure



Problems in 1CLICK-1 X-string-Nugget Matching

- Granularity problem

Phone: 046-223-3636. Fax: 046-223-3630.
Address: 243-8551.
Email: soumu@shonan

X-string

Match?

- hone number: 046-223-3636
- Fax number: 046-223-3630
- Address: 118-1 Nurumizu, Atsugi, 243-8551

Nuggets

- Importance problem

Nugget

Very famous

Less famous

Her main work includes “I Will Always Love You”, "Lover for Life”, and “How Will I Know”.

What’s the score?

Breaking Nuggets into Finer-grained Units

NEW

Nuggets

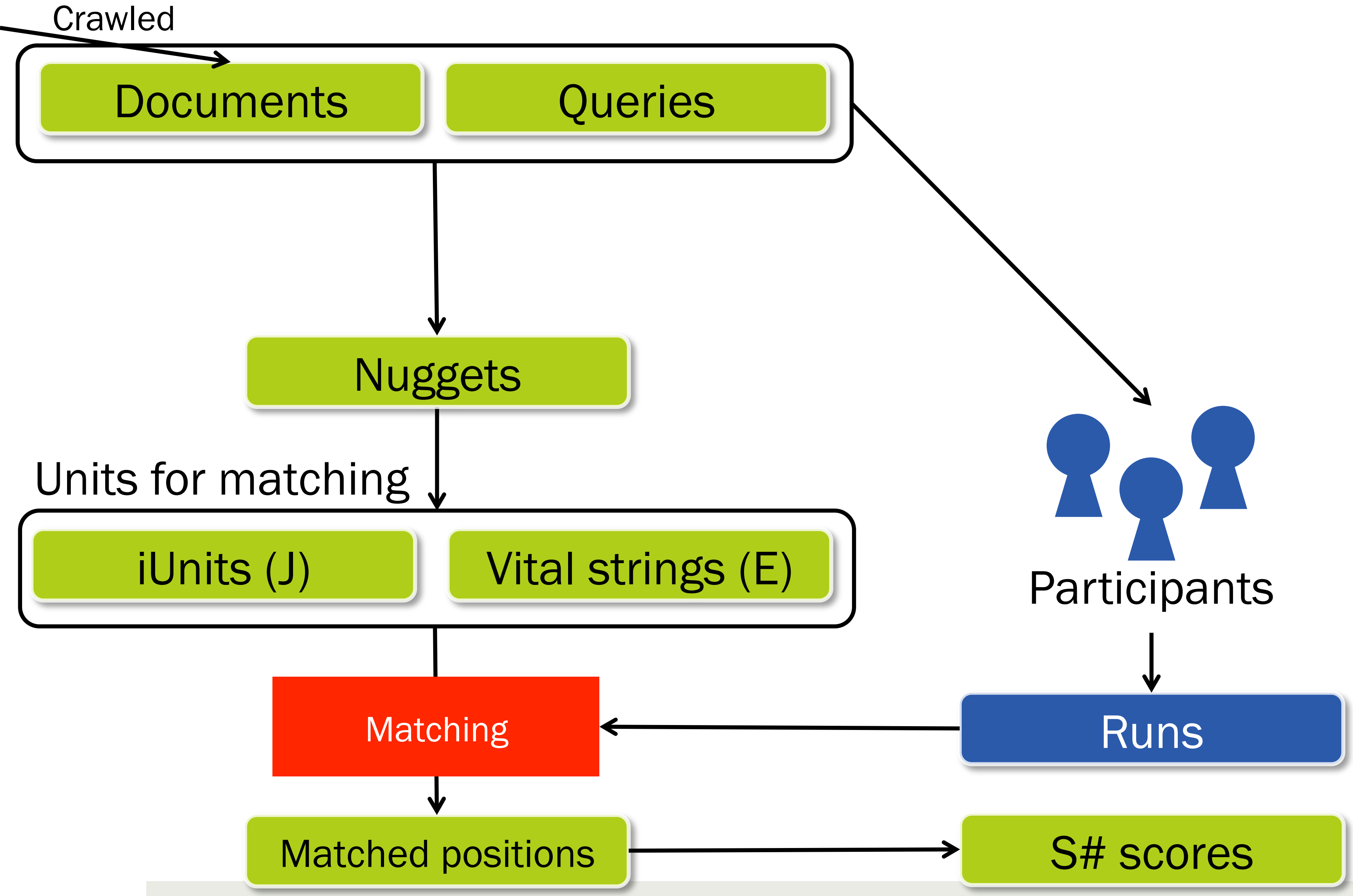
"Murray tried to revive Jackson for ten minutes, at which point he realized he needed to call for help."

Relevant, *Atomic*, and *Dependent* Information pieces

Minimally adequate natural language expression

ID	iUnits	vital string	w	dep
1	Murray tried to revive Jackson	Murray tried to revive	3	
2	Murray tried to revive Jackson for ten minutes	ten minutes	4	1
3	Murray realized he needed to call help	realized he needed to call help	1	1

1CLICK-2 Task Structure



Semi-automatic Matching

NEW

Prev NUIR-E-D-MAND-7 Next Query: marvin gaye influence Category: ARTIST

Instructions

Summary

Automatic matching

After a year as a s... Gaye ranked as the label's top-selling solo artist during the sixties. Due to solo hits including "How Sweet It Is (To Be Loved By You)", "Ain't That Peculiar", "I Heard It Through the Grapevine" and his duet singles with singers such as Mary Wells and Tammi Terrell, he was crowned "The Prince of Motown" and "The Prince of Soul". Notable for fighting the hit-making but restrictive Motown process in which performers and songwriters and producers were kept separate, Gaye proved with albums like his 1971 What's Going On and his 1973 Let's Get It On that he was able to produce music without ... Motown ... Jackson ... ing the Let's G... and I Want You albums helped influence the quiet storm, urban adult contemporary and slow jam genres.

Manual matching

Vital Strings

Search:

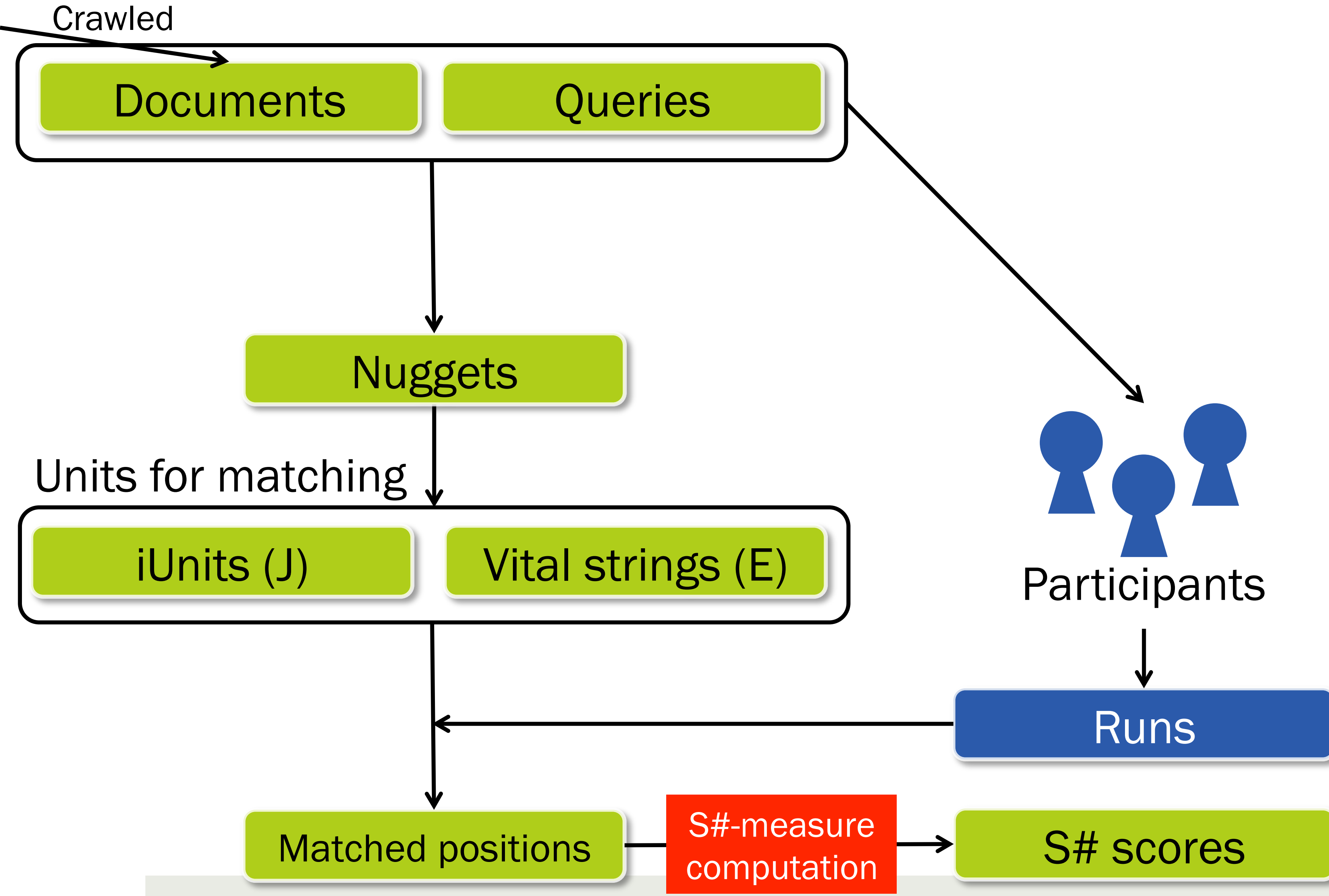
Vital String	Context	Start	End	Importance
10. Let's Get It On		570	586	Low
11. inspiring fellow Motown artists		651	683	Low
12. Stevie Wonder and Michael Jackson		691	725	Low
13. mid-1970s work influenced quiet storm		745	846	High
14. mid-1970s work influenced slow jam genres		745	893	High
15. mid-1970s work influenced urban adult		745	875	High

Clear Position Save Changes Delete Vital String

New Vital String: Vital String Text

Dependencies (e.g. 1,12) Add Vital String

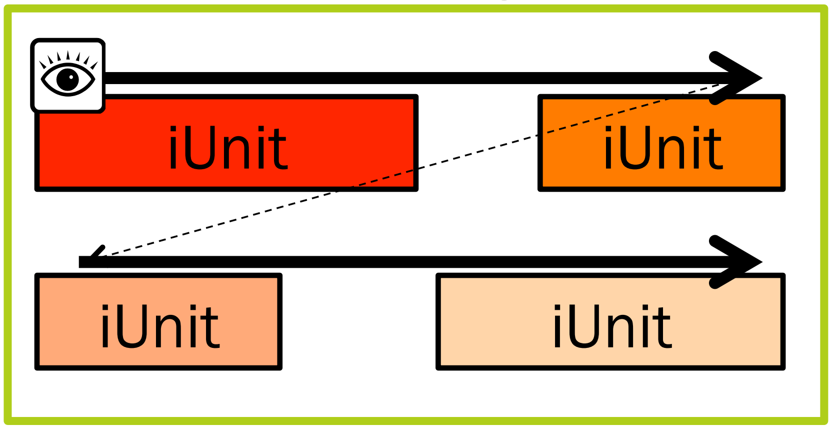
1CLICK-2 Task Structure



S, T, and S#-measure

S-measure

X-string



$$S = \frac{1}{Z} \sum_{i \in M} w(i) d(i)$$

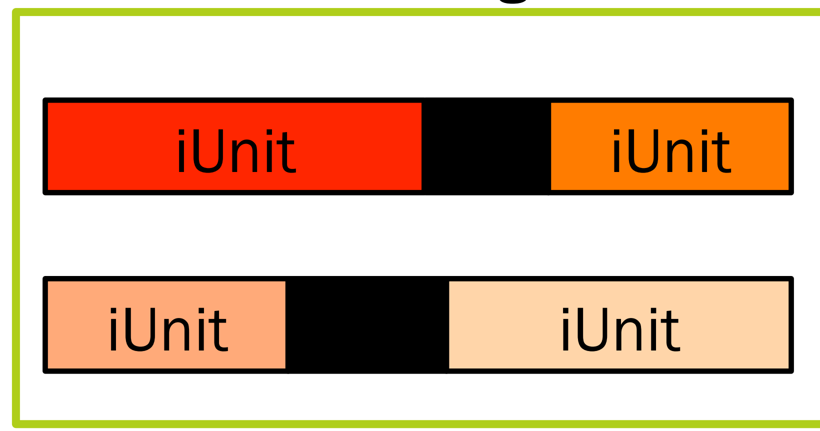
w: weight, Z: normalization factor

$$d(i) = \max(0, L - offset(i))$$

discounts the iUnit (or VS) weight based on its offset

T-measure NEW

X-string

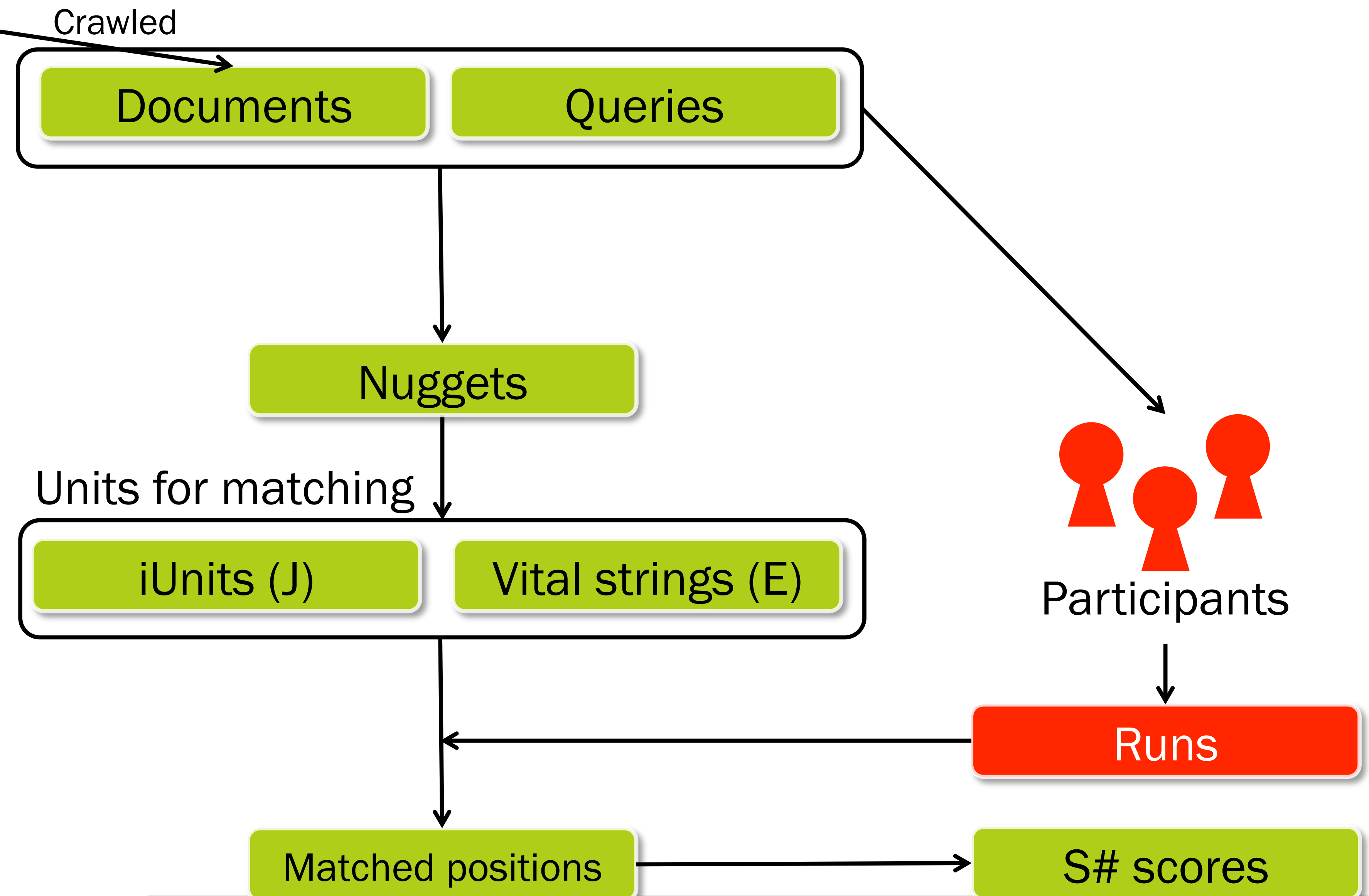


T = % of matched text

S#-measure

The harmonic mean of S and T (official evaluation metric in 1CLICK-2)

1CLICK-2 Task Structure



Participants

English (5 teams, 28 runs)

team name	MAIN	QC	organization
KUIDL	4	2	Kyoto University
NSTDB	6	0	Nara Institute of Science and Technology
NUIR	8	0	Northeastern University, USA
udem	4	0	University of Montreal
ut	2	2	University of Twente

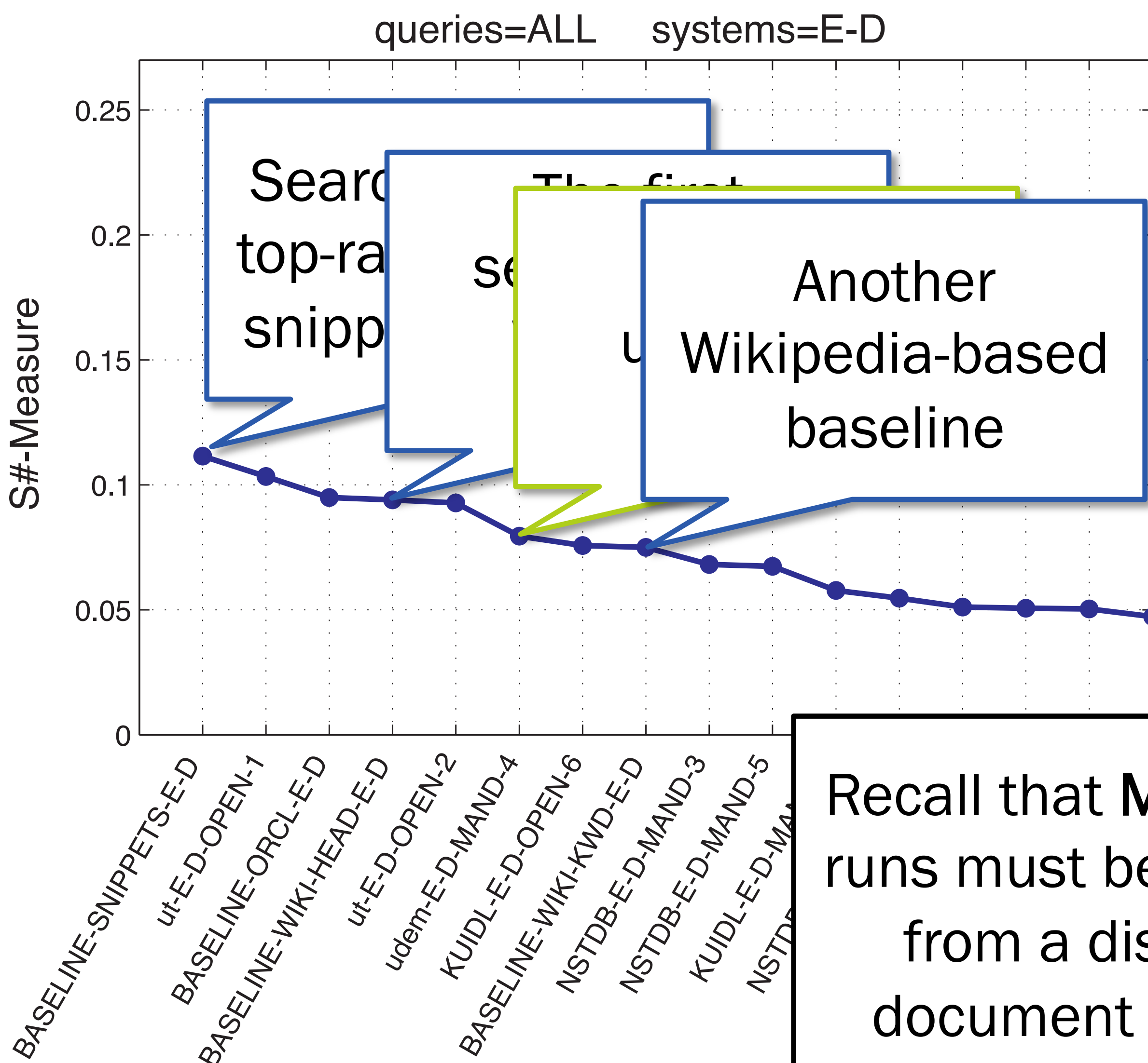
Japanese (5 teams, 22 runs)

team name	MAIN	QC	organization
HUKB	0	2	Hokkaido University
KUIDL	4	2	Kyoto University
MSRA	4	1	Microsoft Research Asia
NUTKS	0	6	Nagaoka University of Technology
TTOKU	3	0	Tokyo Institute of Technology

Results

DESKTOP (1,000 chars) runs in English 1CLICK-2

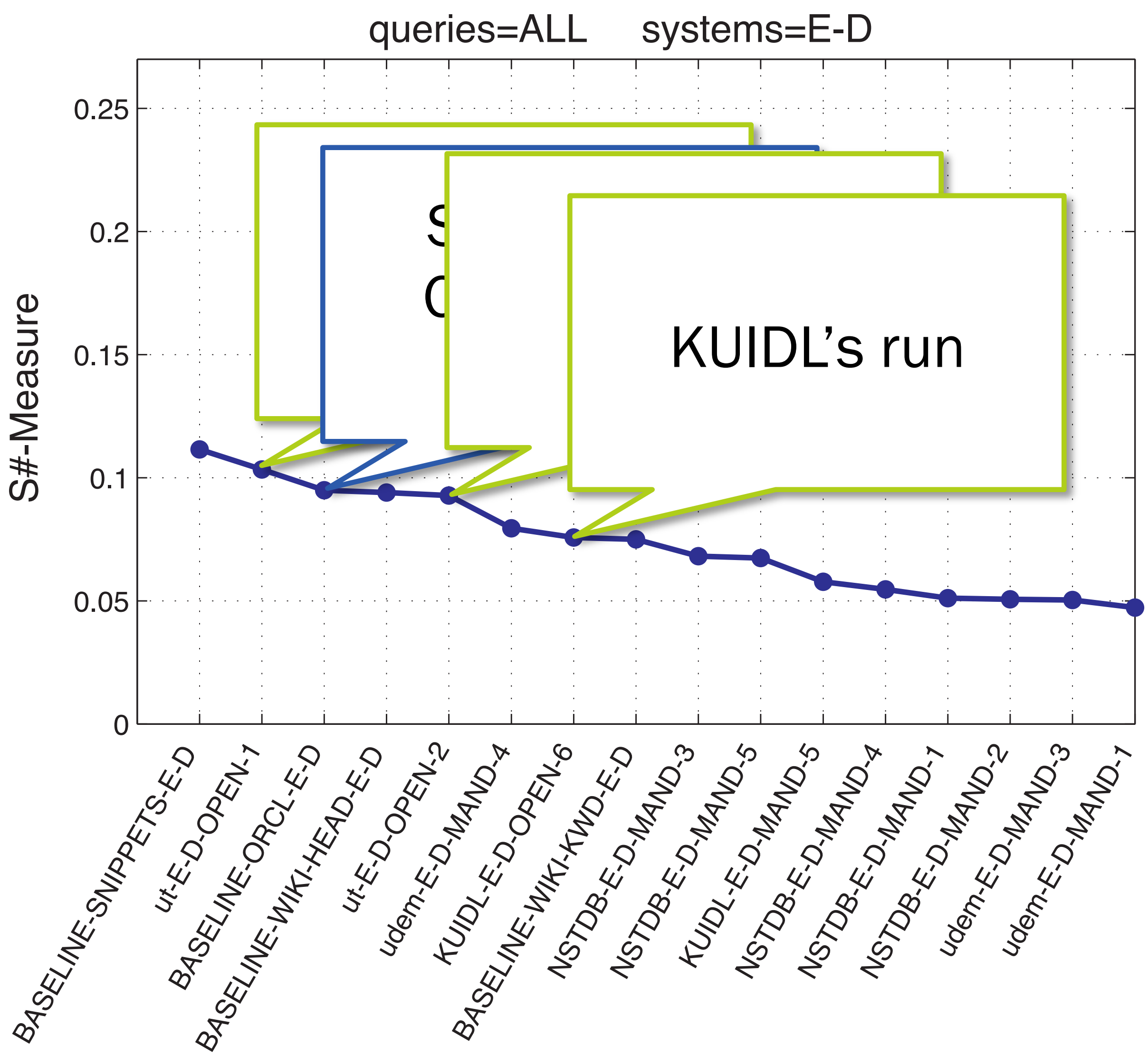
E



Recall that **MANDATORY** runs must be generated from a distributed document collection

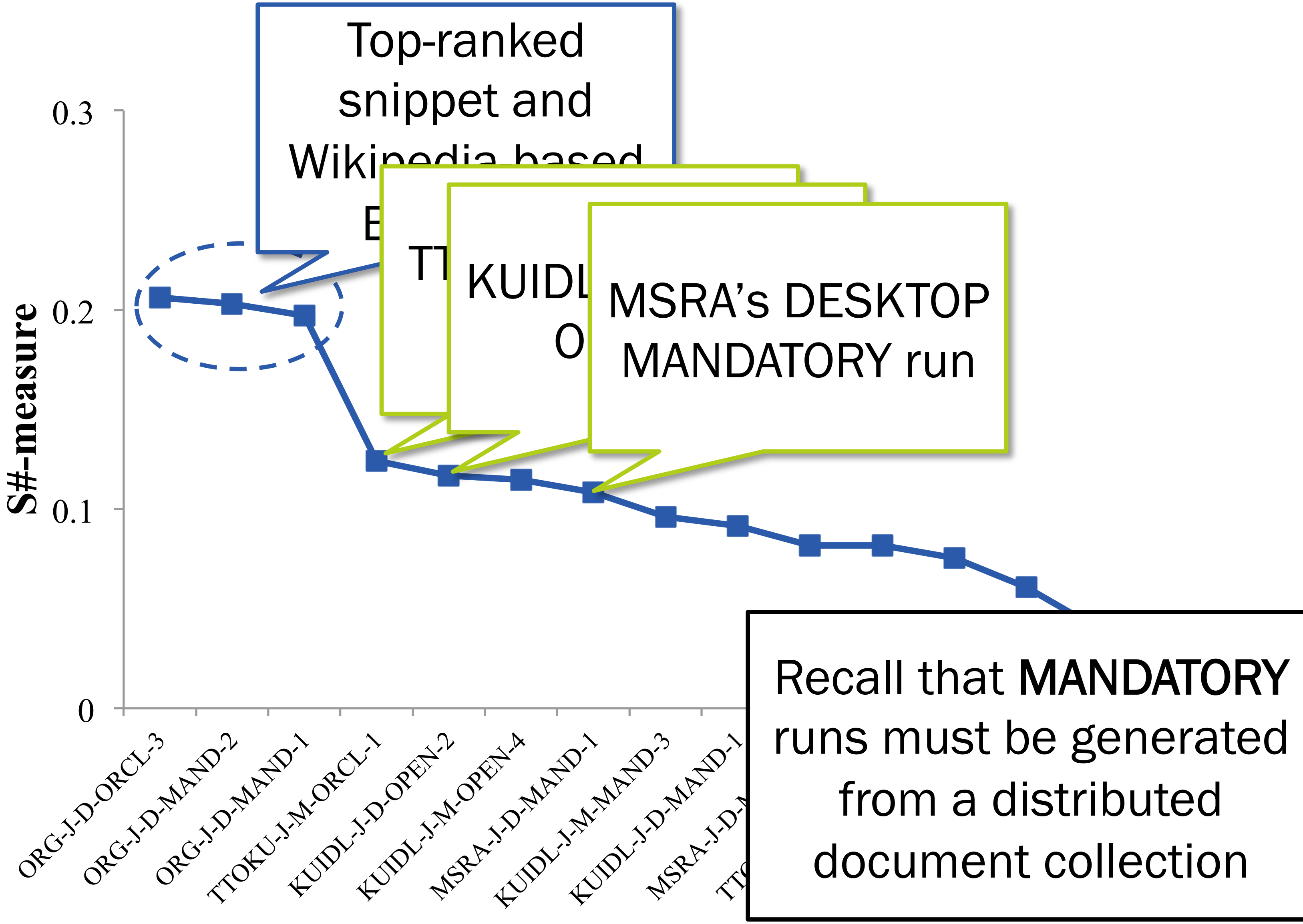
DESKTOP (1,000 chars) runs in English 1CLICK-2

E



Runs in Japanese 1CLICK-2

J

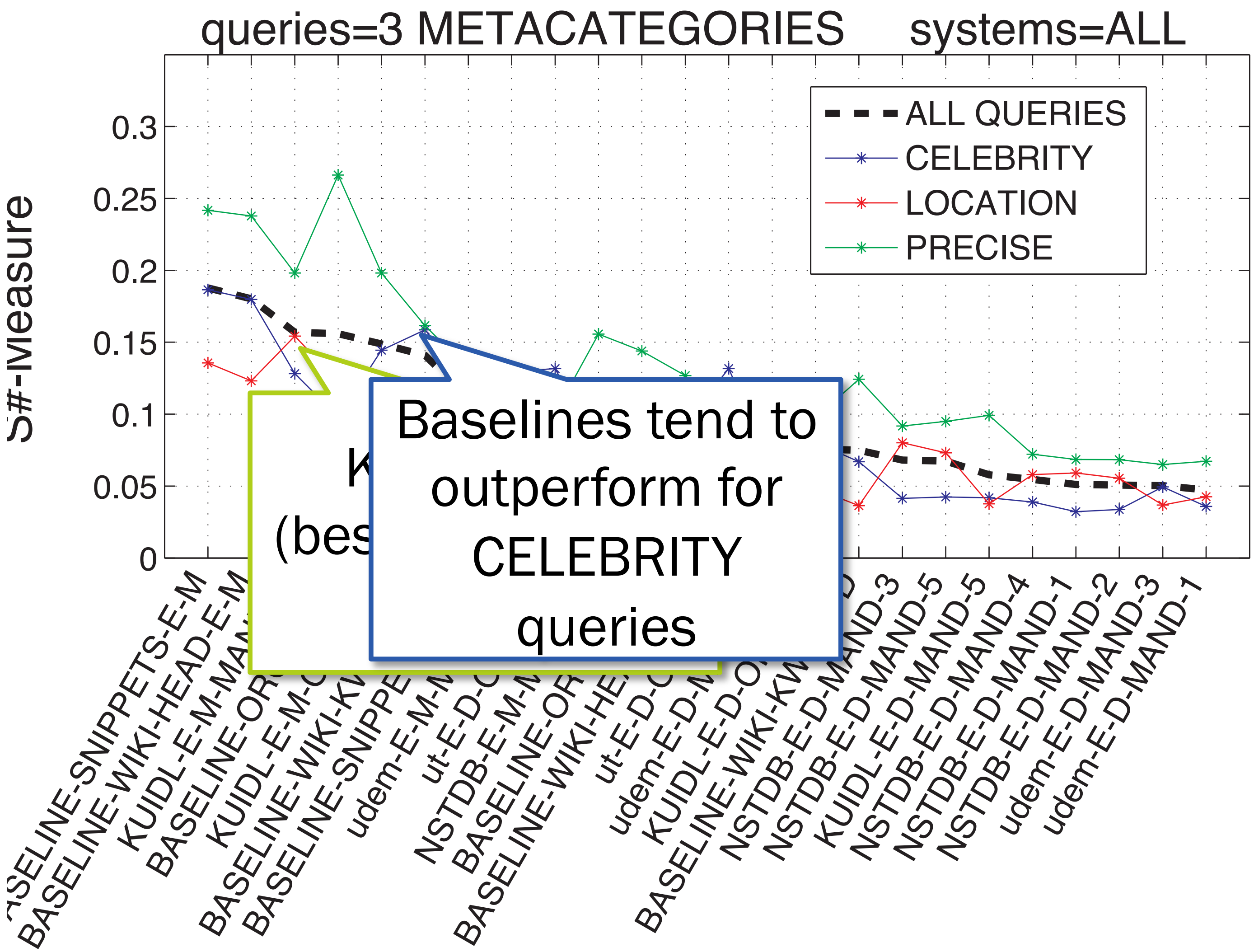


Recall that **MANDATORY** runs must be generated from a distributed document collection

Too Strong Baselines?

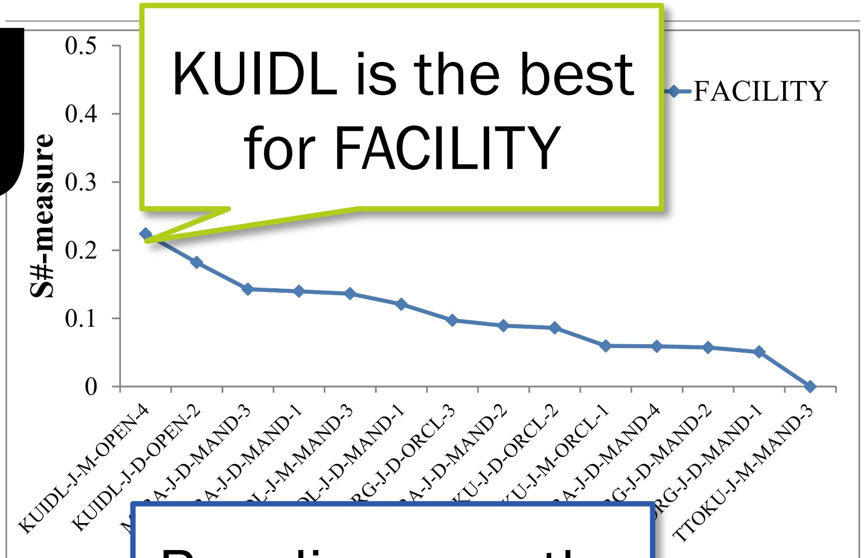
Per-query-type Analysis in English 1CLICK-2

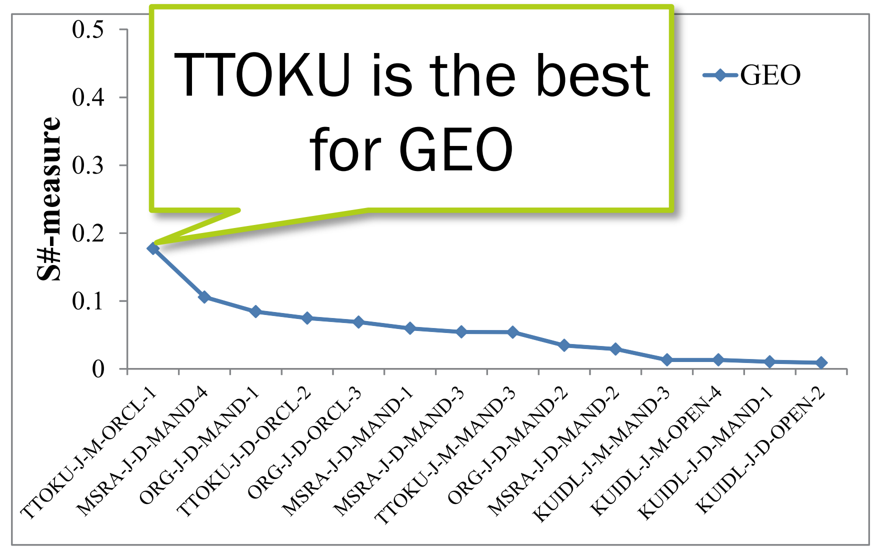
E

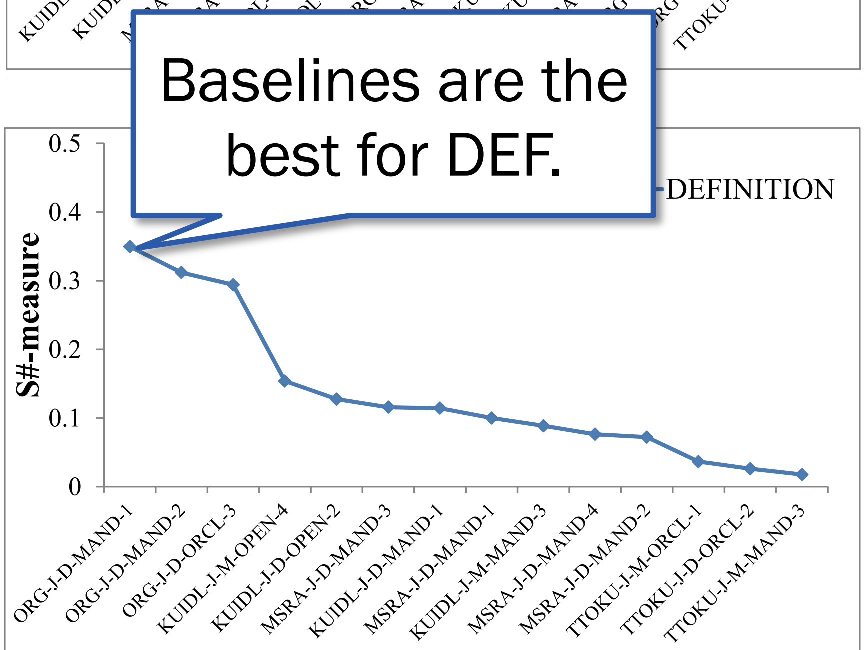


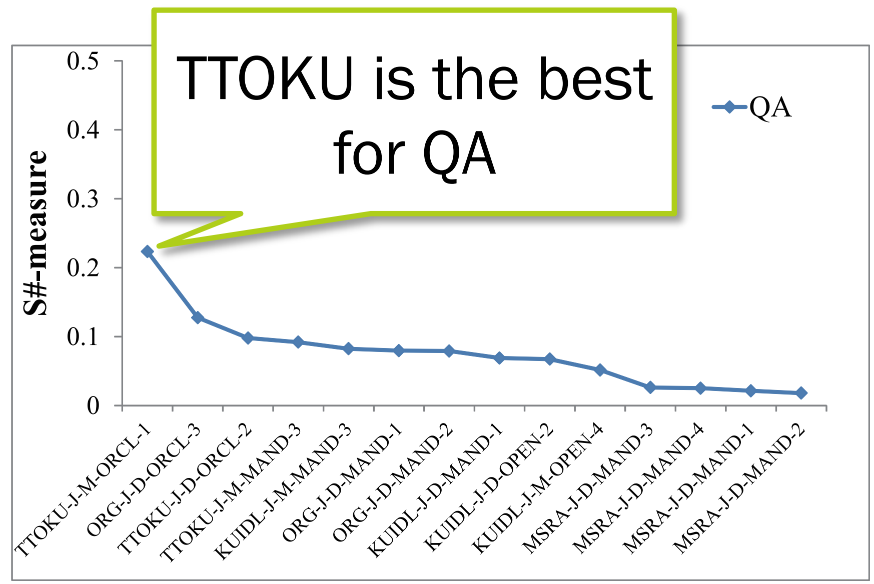
Per-query-type Analysis in Japanese 1CLICK-2

J









Summary of Findings

- Overall
  - Baselines outperformed participants' runs
  - <u>ut is the second best in English 1CLICK-2</u>
- ARTIST, ACTOR, POLITICIAN, and ATHLETE
  - <u>Top performer: Baselines</u>
- FACILITY and LOCATION
  - <u>Top performer: KUIDL and TTOKU</u>
- DEFINITION and QA
  - <u>Top performer: TTOKU</u>

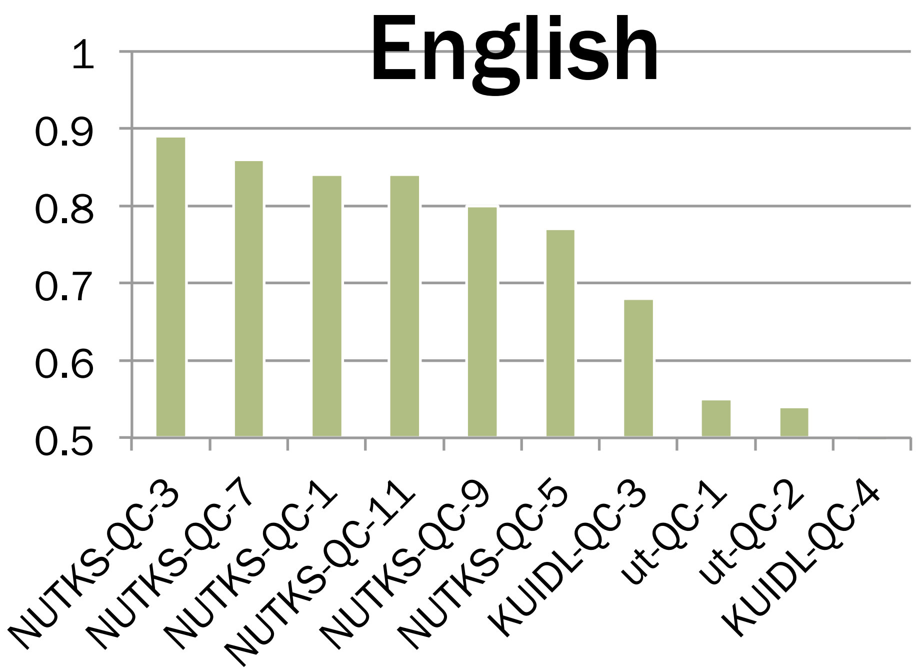
Possible Problems

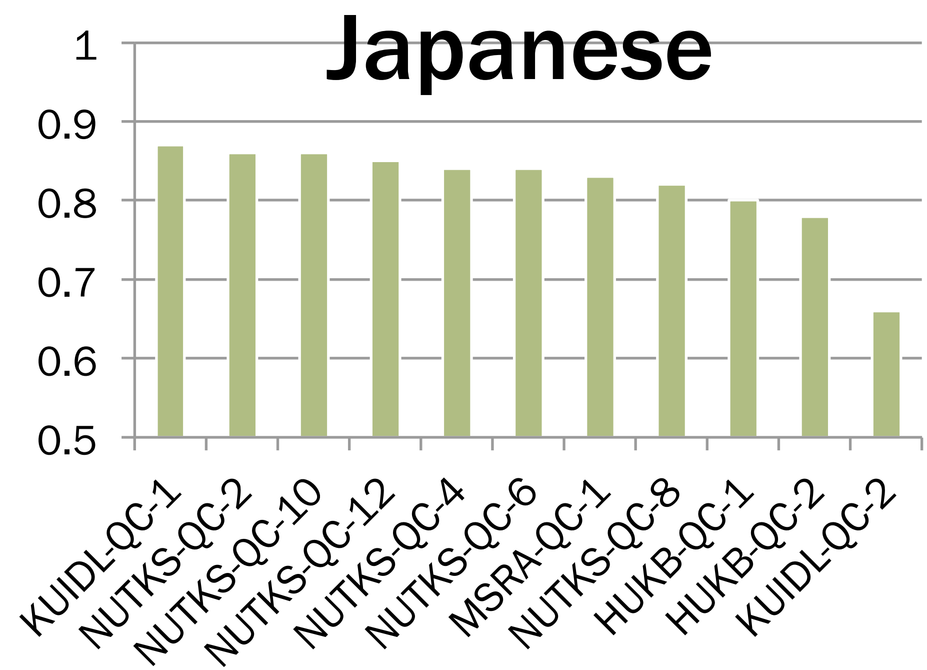
- Readability problem
  - Assessor matching mistakes are more probable on crabbed X-strings than readable ones (e.g. our simple baselines)

["Year": "Producer(s)*"], ["Instruments": "Vocals, piano"], ["1996": "Jagged Little Pill"], ["1998": "Daniel Lanois"], ["Title":	VS	Life and career 1963–1976: Early life Whitney Houston was born in what was then a middle-income neighborhood in

- Wikipedia-is-enough problem
  - For single-term queries, the first sentences from a Wikipedia article are effective enough
  - While specified queries such as "michael jackson death" require a summary from multiple documents

Query Classification Subtask Results





- 0.85+ accuracy achieved (by NUTKS&KUIDL)
  - DIFFICULT: DEFINITION type
  - EASY: CELEBRITY types (ARTIST, ACTOR, etc.)

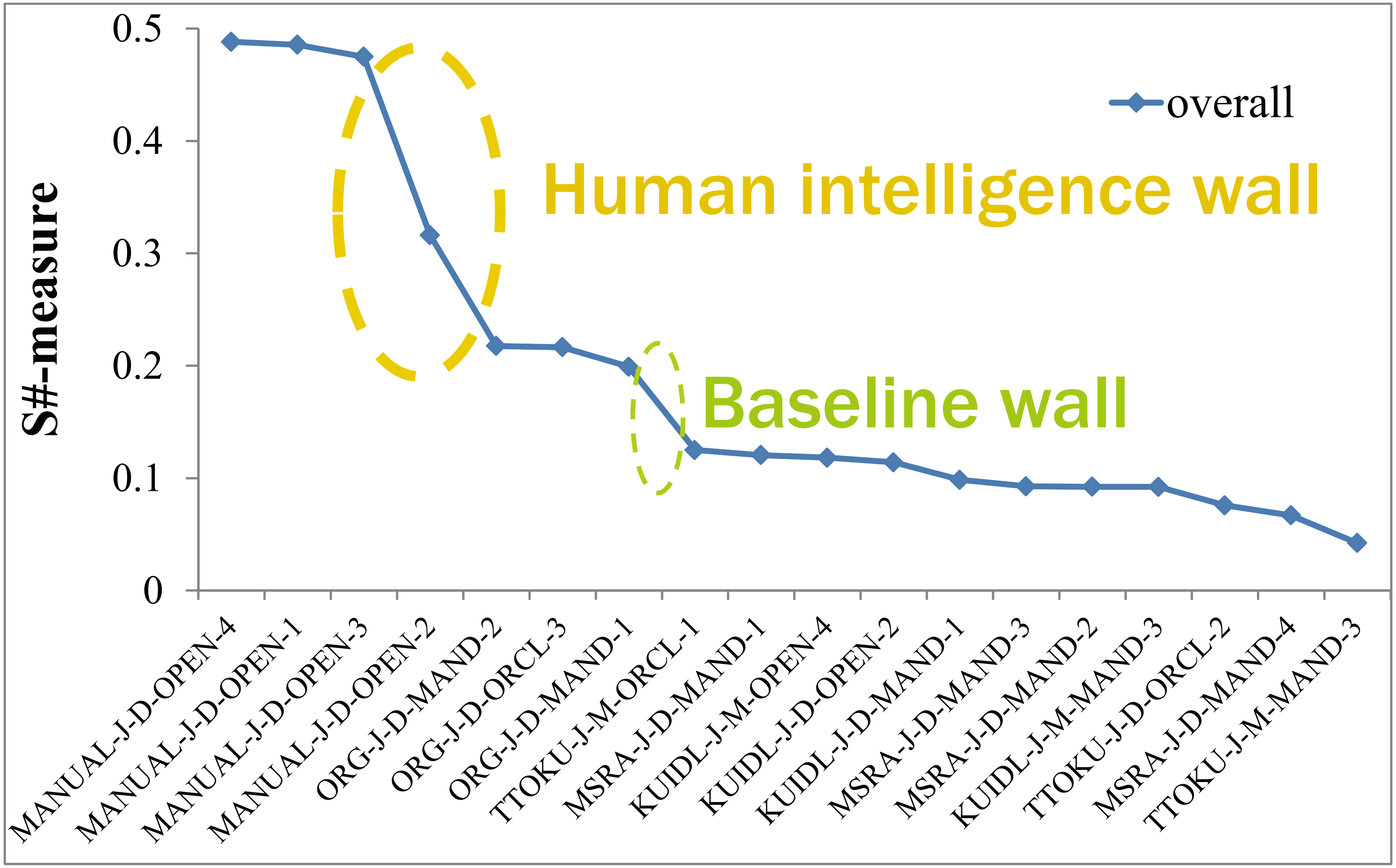
Summary and Future work

Summary

- 1CLICK is immediate and direct information access that focuses more on information retrieval
- Several new features in 1CLICK-2
  - A new subtask
  - Semi-automatic nugget extraction
  - Finer-grained units for matching
  - Semi-automatic matching between X-strings and VSs
- Results
  - Opportunity for big improvement
  - Some runs show good performances for some query types
  - Readability problem for both participants and organizers

The NEXT 1CLICK

JUMP THE TWO WALLS

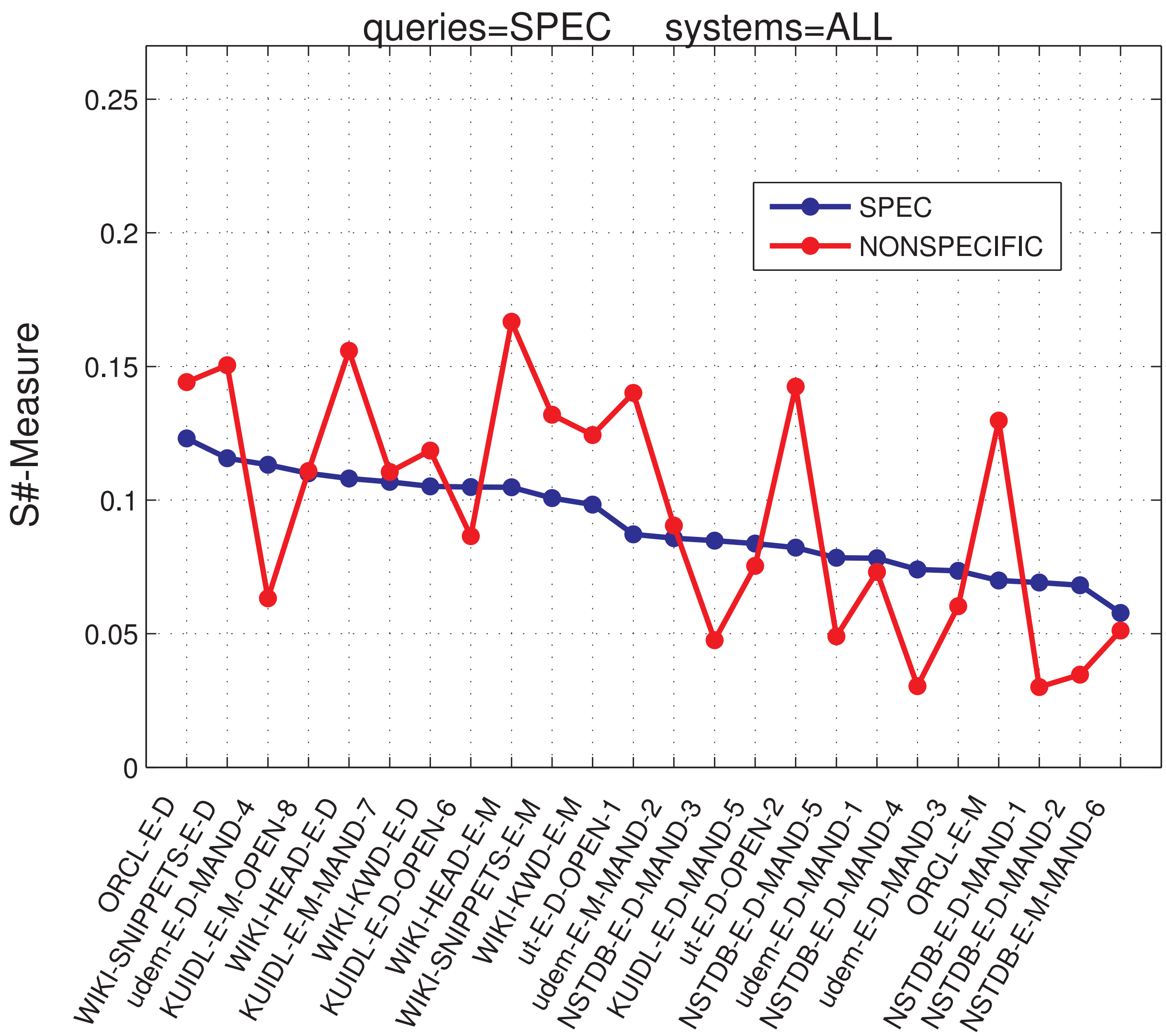


Welcome You to 1CLICK Session on DAY-4

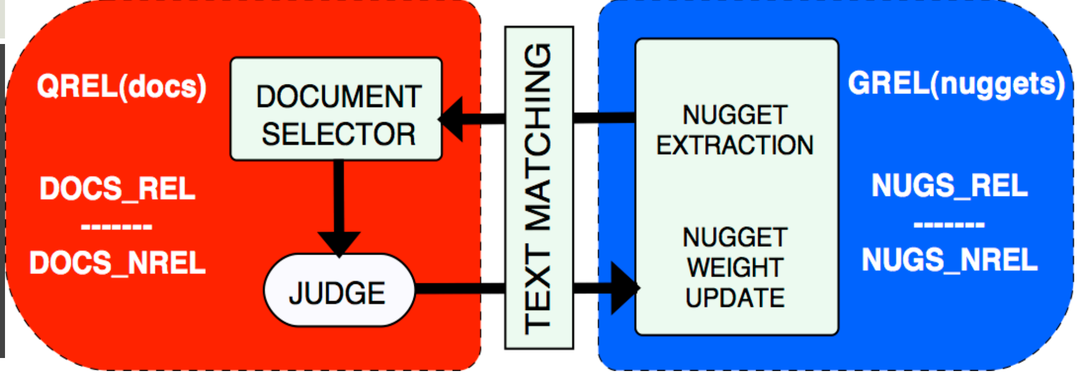
- *TTOKU Summarization Based Systems at NTCIR-10 1CLICK-2 task*
  - Tokyo Institute of Technology team: impressive QA performance
- *MSRA at NTCIR-10 1CLICK-2*
  - Microsoft Research Asia team:
    top performer among Japanese MANDATORY runs
- *An API-based Search System for One Click Access to Information*
  - University of Twente team: top performer in English 1CLICK-2
- *Hunter Gatherer: UdeM at 1CLICK-2*
  - Université de Montréal team:
    top performer among English MANDATORY runs
- *XML Element Retrieval@1CLICK-2*
  - Nara Institute of Science and Technology team:
    unique approach to 1CLICK

Thank you!

Specific vs. Unspecific



Semi-automatic Nugget Extraction



- Iterative reinforcement between nuggets/documents
  - Document(s) selected based on matching high quallity nuggets
  - Document(s) assessed for binary relevance Rel/NRel
  - New Nuggets introduced = sentences from Rel documents
  - Nuggets updated quality
    - Rel docs matching a nugget increase nugget quality
    - NRel docs matching a nugget decrease nugget quality
    - Non-matching docs don't affect nugget quality
  - Judge can review/assess nuggets directly
  - Judge can extract manually a nugget (different than a sentence from Rel docs)

Readability Scores

ble 16: Japanese Subtask: Mean of the sum of two assessors' readability and trustworthiness score

run name	readability	trustworthiness
KUIDL-J-D-MAND-1.tsv	-1.3	-1.48
KUIDL-J-D-OPEN-2.tsv	-1.32	-1.27
KUIDL-J-M-MAND-3.tsv	-1.23	-2.42
KUIDL-J-M-OPEN-4.tsv	-1.67	-2.46
MANUAL-J-D-OPEN-1.tsv	0.92	1.08
MANUAL-J-D-OPEN-2.tsv	0.54	0.74
MANUAL-J-D-OPEN-3.tsv	0.92	0.96
MANUAL-J-D-OPEN-4.tsv	0.57	0.91
MSRA-J-D-MAND-1.tsv	-1.43	-1.47
MSRA-J-D-MAND-2.tsv	-1.8	-1.97
MSRA-J-D-MAND-3.tsv	-1.89	-2.08
MSRA-J-D-MAND-4.tsv	-2.14	-2.16
ORG-J-D-MAND-1.tsv	-0.26	-0.13
ORG-J-D-MAND-2.tsv	1.5	0.82
ORG-J-D-ORCL-3.tsv	-0.31	0.09
TTOKU-J-D-ORCL-2.tsv	-2.35	-2.27
TTOKU-J-M-MAND-3.tsv	-2.68	-3.23
TTOKU-J-M-ORCL-1.tsv	-1.55	-2.28

Significant Test Result

Table 12: Japanese Subtask: p-values of two-sided randomized Tukey's HSD in terms of S$\sharp$-measure performances over 100 Japanese queries ($L = 500$). Bold font indicates p-values $< \alpha = 0.05$.

		KUIDL				MSRA				ORG			TTOKU		
		1	2	3	4	1	2	3	4	1	2	3	1	2	3
KUIDL	1		0.997	1.000	0.998	1.000	1.000	1.000	0.969	**0.000**	**0.000**	**0.000**	0.952	1.000	0.378
	2			1.000	1.000	1.000	0.749	0.909	0.254	**0.007**	**0.002**	**0.001**	1.000	0.909	**0.010**
	3				1.000	1.000	1.000	1.000	0.910	**0.000**	**0.000**	**0.000**	0.987	1.000	0.245
	4					1.000	0.810	0.943	0.311	**0.005**	**0.001**	**0.001**	1.000	0.942	**0.017**
MSRA	1						0.949	0.994	0.541	**0.001**	**0.000**	**0.000**	1.000	0.994	**0.047**
	2							1.000	1.000	**0.000**	**0.000**	**0.000**	0.491	1.000	0.896
	3								1.000	**0.000**	**0.000**	**0.000**	0.716	1.000	0.729
	4									**0.000**	**0.000**	**0.000**	0.103	1.000	0.999
ORG	1										1.000	1.000	**0.025**	**0.000**	**0.000**
	2											1.000	**0.009**	**0.000**	**0.000**
	3												**0.005**	**0.000**	**0.000**
TTOKU	1													0.711	**0.003**
	2														0.731
	3														

Disagreement across Assessors

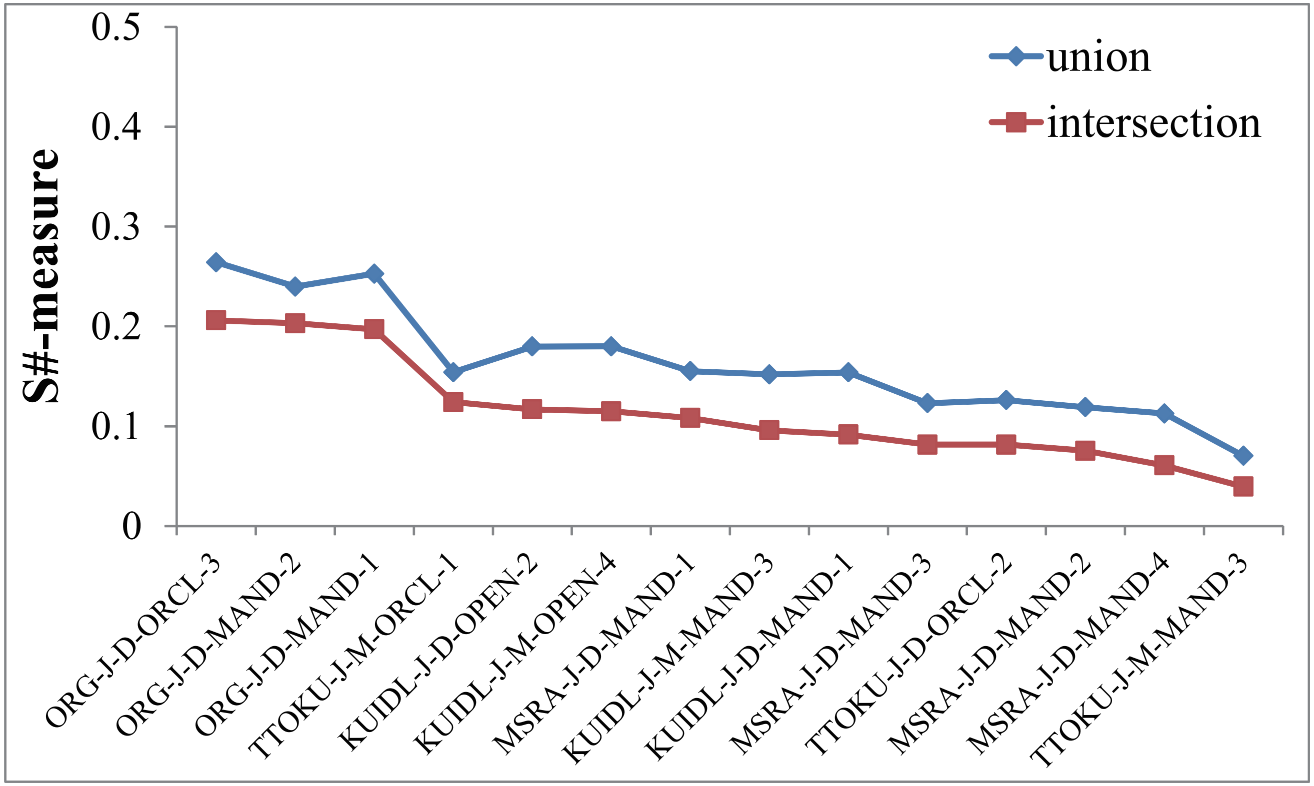


Figure 9: Japanese Subtask: Mean S♯-measure performances over 100 queries ($L = 500$). The x axis represents runs sorted by Mean S♯ with the intersection iUnit match data.

Inter-rater Agreement

Table 23: Japanese Subtask: Inter-rater agreement in terms of Cohen's kappa coefficient, mean absolute error (MAE), and mean square error (MSE).

assessor pairs		Kappa	MAE	MSE
a_7	a_1	0.782	10.4	2100
a_1	a_2	0.738	15.5	4030
a_2	a_3	0.717	5.32	896
a_3	a_4	0.755	3.87	901
a_4	a_5	0.743	6.01	1480
a_5	a_6	0.751	4.77	744
a_6	a_7	0.688	4.18	1110
a_8	a_9	0.818	4.07	620
a_9	a_{10}	0.770	5.46	599
a_{10}	a_{11}	0.704	6.41	863
a_{11}	a_8	0.753	4.61	731
average		0.747	6.41	1280

Correlation across Utility, ROUGE, and S# with Manual/Automatic Matching

	Utility VS			ROUGE VS		Man VS
CATEG	ROUG	Man	Auto	Man	Auto	Auto
ACTOR	0.68	0.58	0.56	0.35	0.34	0.93
ARTIS	0.59	0.53	0.53	0.24	0.32	0.82
ATHLE	0.49	0.49	0.52	0.20	0.17	0.85
POLIT	0.53	0.42	0.45	0.05	0.04	0.71
GEO	0.38	0.36	0.43	-0.05	0.00	0.70
FACIL	0.54	0.42	0.43	0.34	0.33	0.77
DEFIN	0.48	0.63	0.58	0.37	0.29	0.91
QA	0.67	0.51	0.43	0.39	0.28	0.83
ALLQ	0.59	0.62	0.58	0.26	0.18	0.89

Table 2: Pairwise comparisons of evaluation metrics used for this task against utility as perceived by the assessors. Values taken by averaging scores over each category and comparing induced rankings via Kendall's Tau. Man and Auto are manual and automatic matches combined with the $S\sharp$ evaluation metric.

Correlation across Utility, Estimated Readability, and Readability

	Utility VS		Readability VS	
CATEG	Read-F-$S\sharp$	Read-S-$S\sharp$	Read-F	Read-S
ACTOR	0.56	0.58	0.18	0.22
ARTIS	0.56	0.57	0.03	0.13
ATHLE	0.49	0.49	0.41	0.42
POLIT	0.31	0.33	-0.09	-0.13
GEO	0.28	0.27	0.26	0.27
FACIL	0.50	0.50	0.35	0.41
DEFIN	0.56	0.55	0.14	0.09
QA	0.53	0.50	-0.30	-0.17
ALLQ	0.54	0.53	0.16	0.15

Table 3: Kendall's Tau pairwise comparisons of $S\sharp$ using readability metrics versus Utility, and of only the readability scores versus assesed Readability, all using rankings induced by average scores over query categories.

iUnits

- Information pieces that satisfy the following properties
  - Relevant: can satisfy the user's information need
  - Atomic: cannot be broken down into multiple iUnits
  - Dependent: can depend on other iUnits
- Example nugget:
  "Murray tried to revive Jackson for ten minutes, at which point he realized he needed to call for help. "

ID	iUnits	weight	dep
001	Murray tried to revive Jackson	3	
002	Murray tried to revive Jackson for ten minutes	4	001
003	Murray realized he needed to call help	1	001

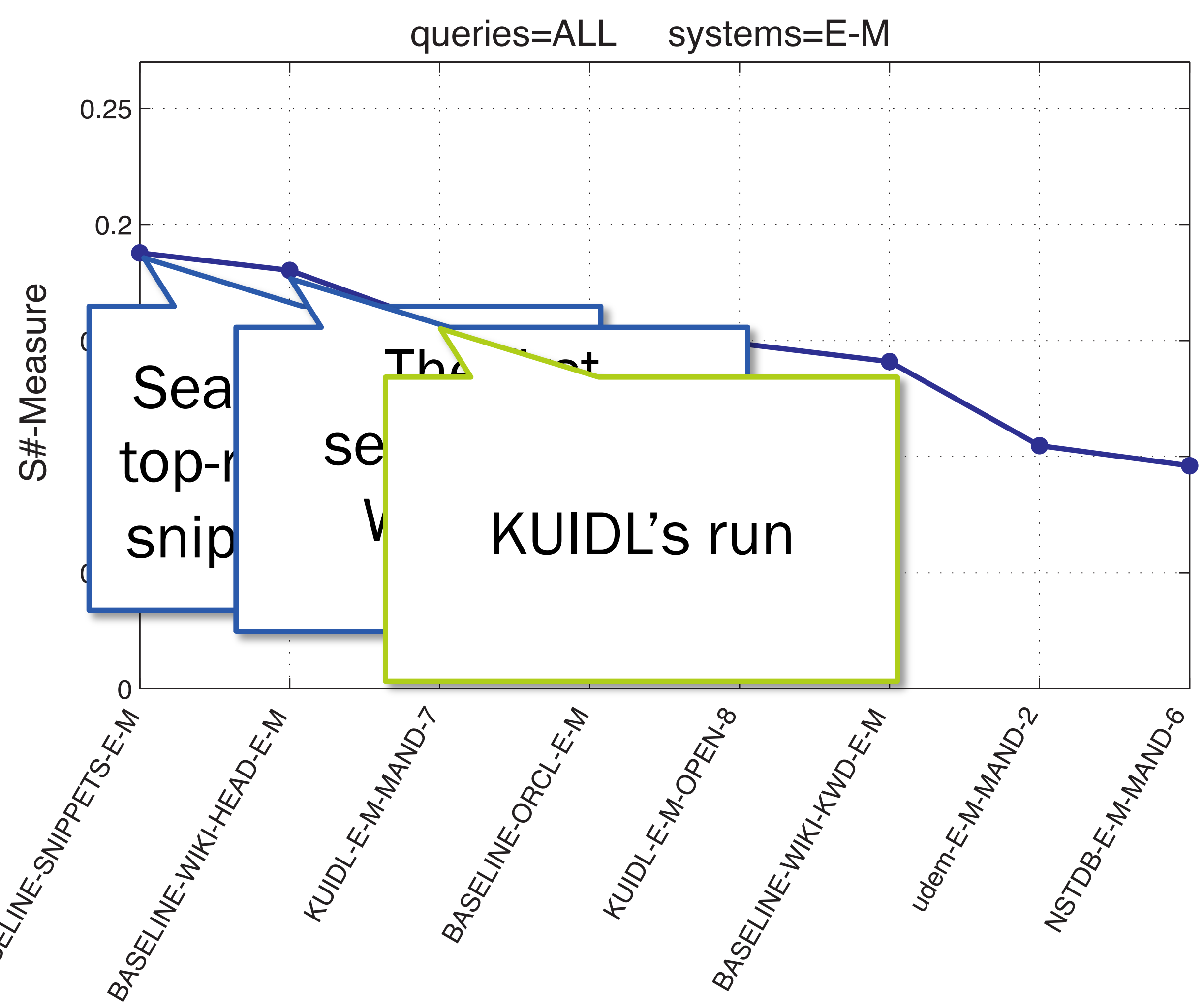
Vital strings

- Minimally adequate natural language expression
  - Obtained from either nuggets or iUnits
- Example nugget:
  "Murray tried to revive Jackson for ten minutes, at which point he realized he needed to call for help. "

ID	iUnits	weight	Dep
001	Murray tried to revive	3	
002	ten minutes	4	001
003	realized he needed to call help	1	001

MOBILE (280 chars) runs in English 1CLICK-2

E

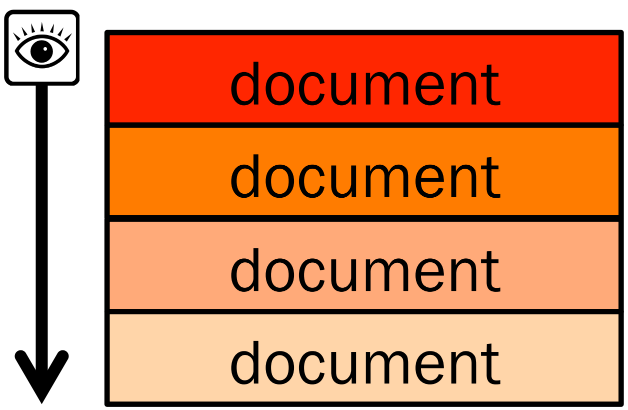


Evaluation Metric: S-measure

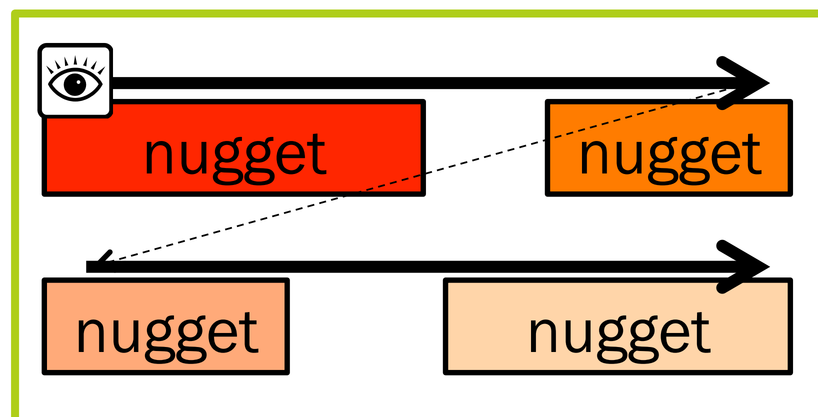
nDCG discounts documents based on ranks

S-measure discounts nuggets based on offsets (positions in *X*-string)

Ranked list of documents

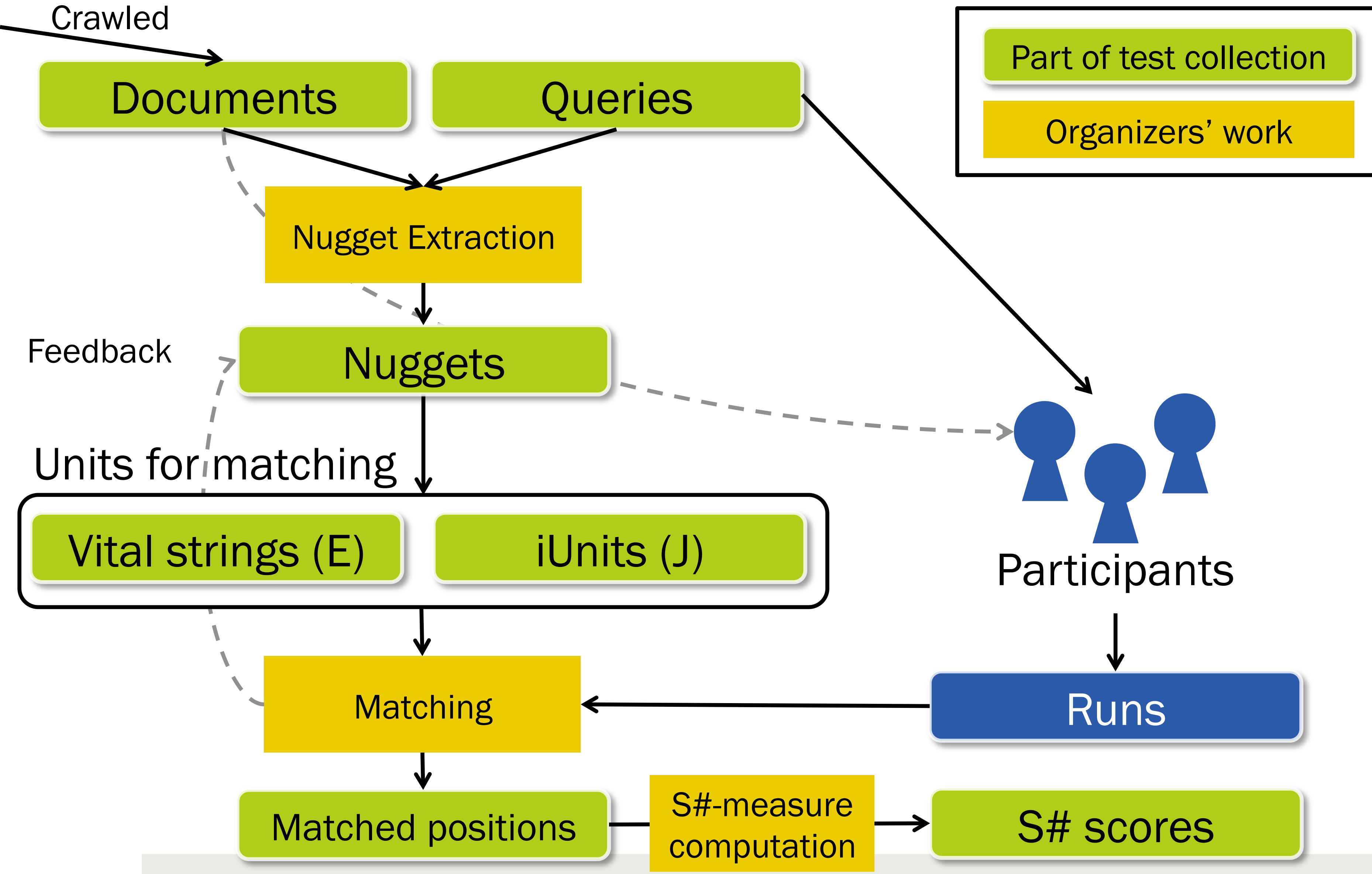


X-string



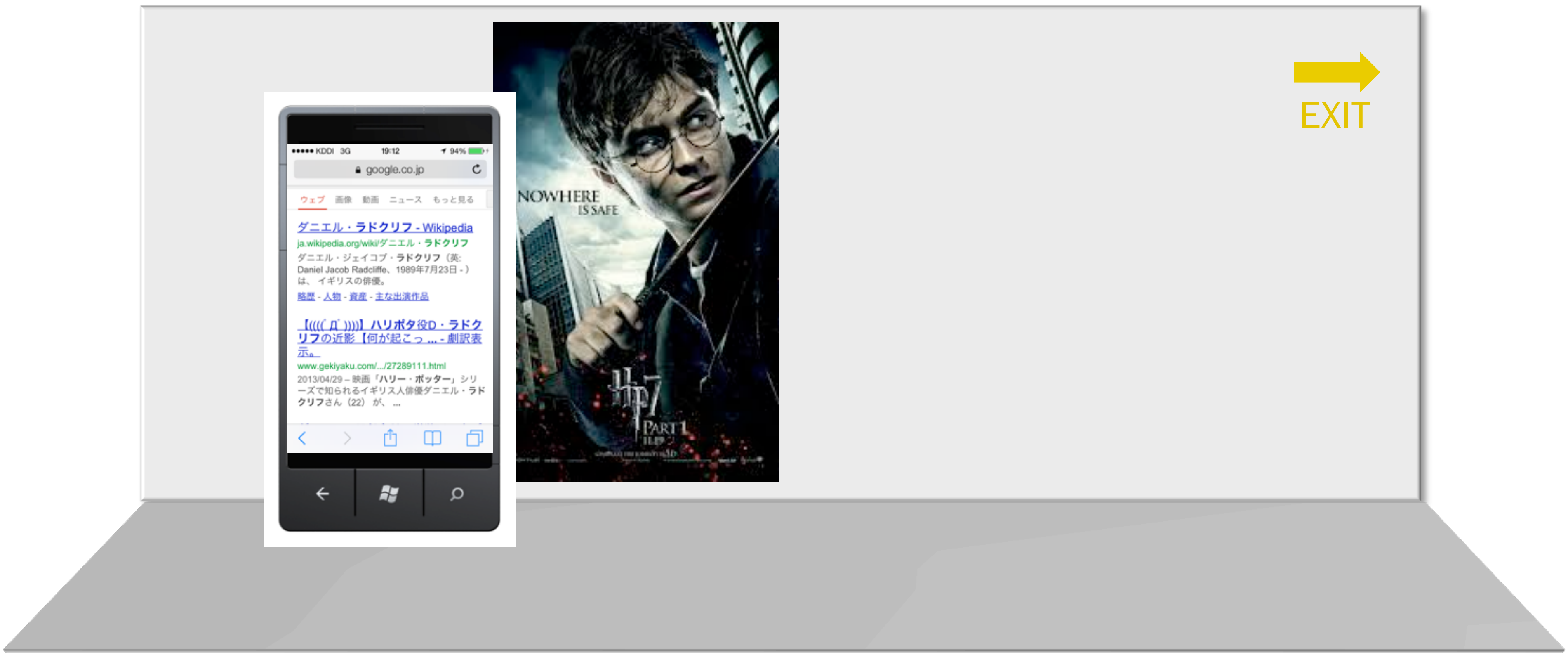
Sakai, Kato, Song: Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access, ACM CIKM 2011

1CLICK-2 Task Structure

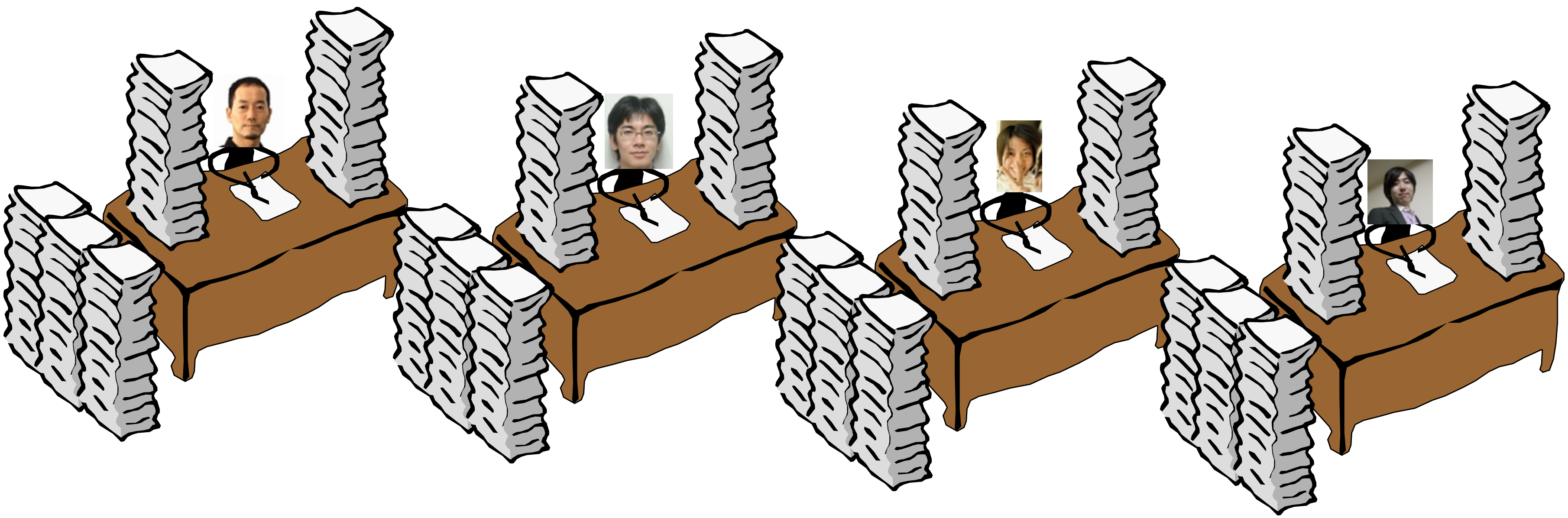


Suppose that …

- Searching for information about his highlights in a movie with a mobile device



Organizers worked hard, too



3,927 nuggets for 100 queries

Very time-consuming process