

Overview of Patent Retrieval Task at NTCIR-4

Atsushi Fujii (Univ. of Tsukuba)

Makoto Iwayama (Hitaci, Ltd.)

Noriko Kando (National Inst. of Informatics)

Introduction

- Large test collections for Human Language Technology (HLT) have been produced in TREC, CLEF, NTCIR
 - Targets are newspaper, technical paper, Web
- Commercial patent retrieval systems have operated for a long time
- But, less attention in the HLT research community

NTCIR-3 Workshop (2001-2002)

- In NTCIR-3, the first effort was made to produce test collection for patent IR
 - technology survey
 - requested to search for patents related to a specific technology (e.g., gasoline direct-injection engine)
- But, process of patent IR differs depending on the purpose
 - technology survey, invalidity search, etc.
- We performed a different task in NTCIR-4

NTCIR-4 Workshop (2003-2004)

- NTCIR workshop is in one and half years
 - difficult to explore long-term research topics
- Two different patent tasks were performed
 - invalidity search task: short-term
 - patent map generation task: long-term
 - feasibility study (FS) task

focus of today's talk

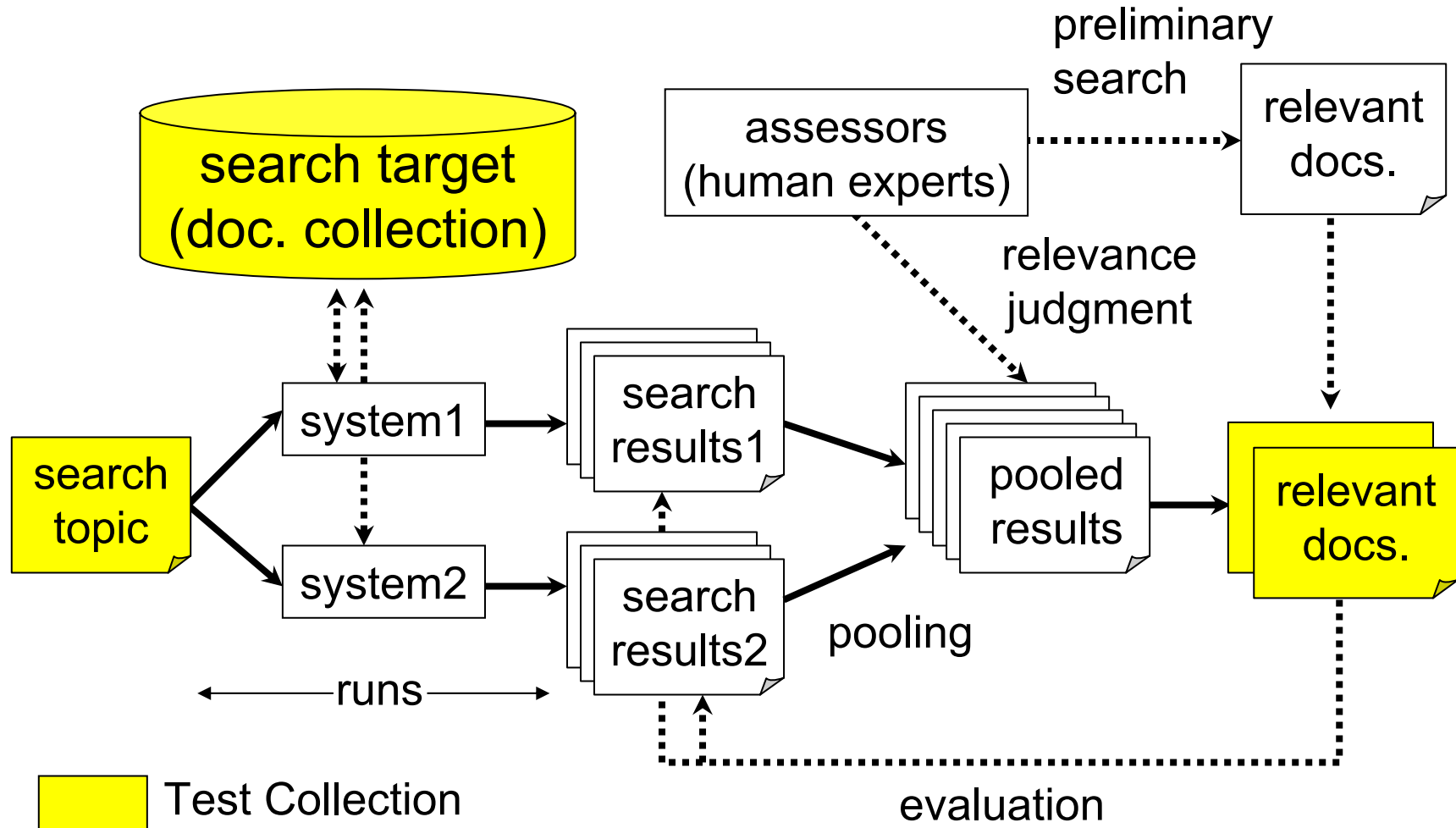
NTCIR-4 Workshop (2003-2004)

- NTCIR workshop is in one and half years
 - difficult to explore long-term research topics
- Two different patent tasks were performed
 - **invalidity search task**: short-term
 - patent map generation task: long-term
 - feasibility study (FS) task

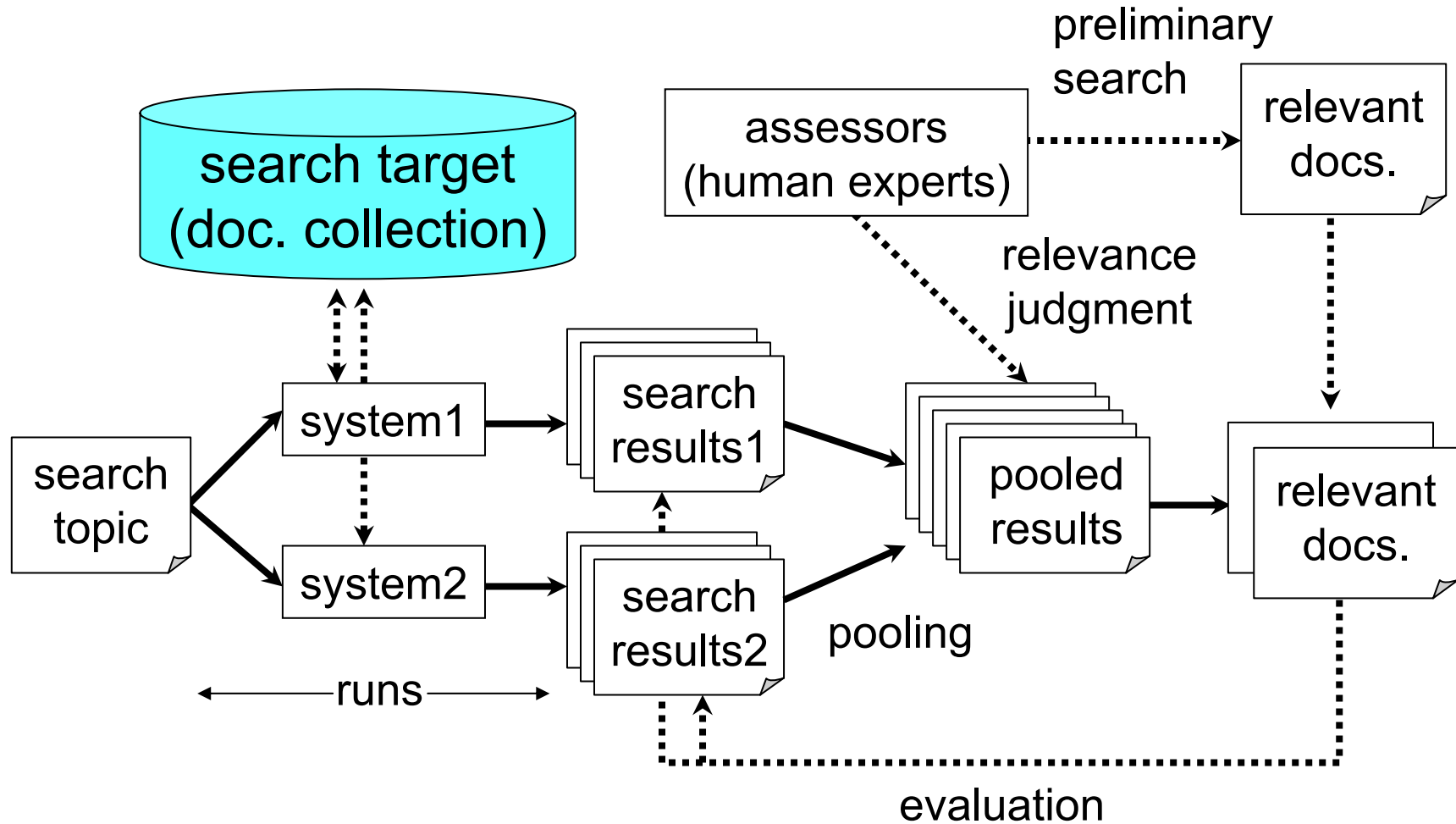
Invalidity search task

- Find the patents that can invalidate the demand in a patent application (claim)
 - given a patent claim, each group searches a collection for patents similar to the claim
- This task is usually performed by
 - examiners in a government patent office
 - searchers of IP division in private companies
- This can be seen as patent-to-patent associative retrieval
 - both queries and documents are patents

Process of producing test collection



Process of producing test collection



Document collection

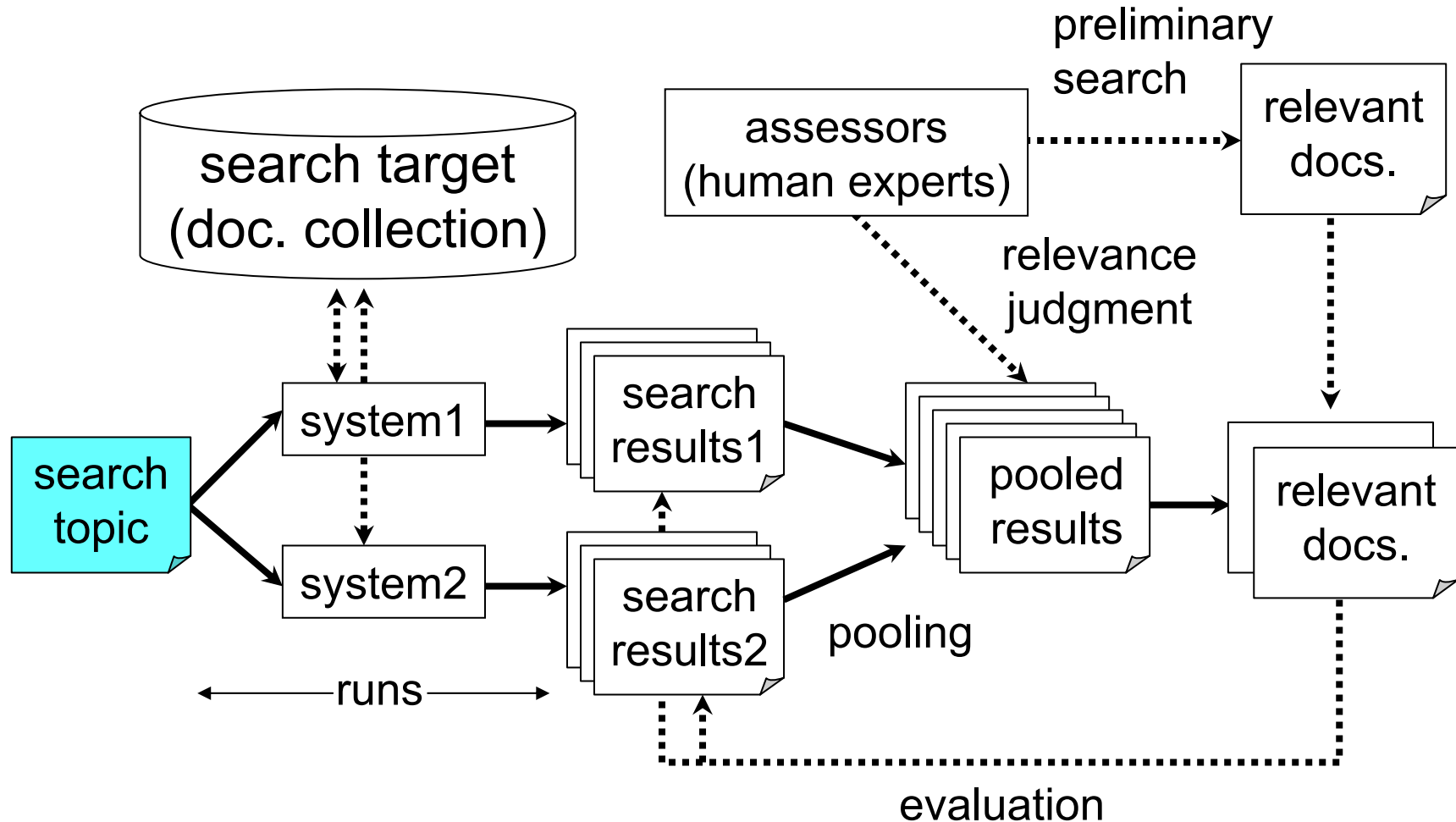
editing

- Unexamined patent application
 - Japanese full text published in 1993-1997
 - 1.7M documents (40GB)
- JAPIO Patent Abstract
 - professional abstracts
 - length is standardized in approx. 400 characters
 - vocabulary is controlled
- Patent Abstracts of Japan (PAJ)
 - English translations of JAPIO Abstract

provided for NTCIR-4

translation

Process of producing test collection



Search topics

- Japanese patent application rejected by Japanese Patent Office (JPO)
 - at least one relevant document exists
- 34 topics were selected by members of “Japan Intellectual Property Association” (JIPA)
 - patent search experts in IP division
 - also in charge of relevance judgment
- English, Korean, and simplified/traditional Chinese translations for cross-language patent IR

Search topics (cont.)

- In preliminary study, the number of relevant documents for a topic was small (< 10)
- Evaluation results obtained with our collection can potentially be unreliable
- QA task overcomes this problem by increasing the number of questions (> 100)
- So, we produced additional topics

Additional topics

- We produced 69 additional topics
- Additional topics are also Japanese patent applications rejected by JPO
- We used only the citations provided by JPO as relevant documents
 - no additional human judgments were needed

Example search topic

Date of filing (May 27, 1996)

<TOPIC>

<NUM>008</NUM>

<LANG>EN</LANG>

<FDATE>19960527</FDATE>

<CLAIM>(Claim 1) A sensor device, characterized in that an open recessed part is formed on a box-shaped forming base, a conductive film of a designated pattern is formed on the surface of the forming base including the inner surface of the recessed part, an element for a sensor is bonded to the recessed part, and the forming base is closed with a cover.</CLAIM>

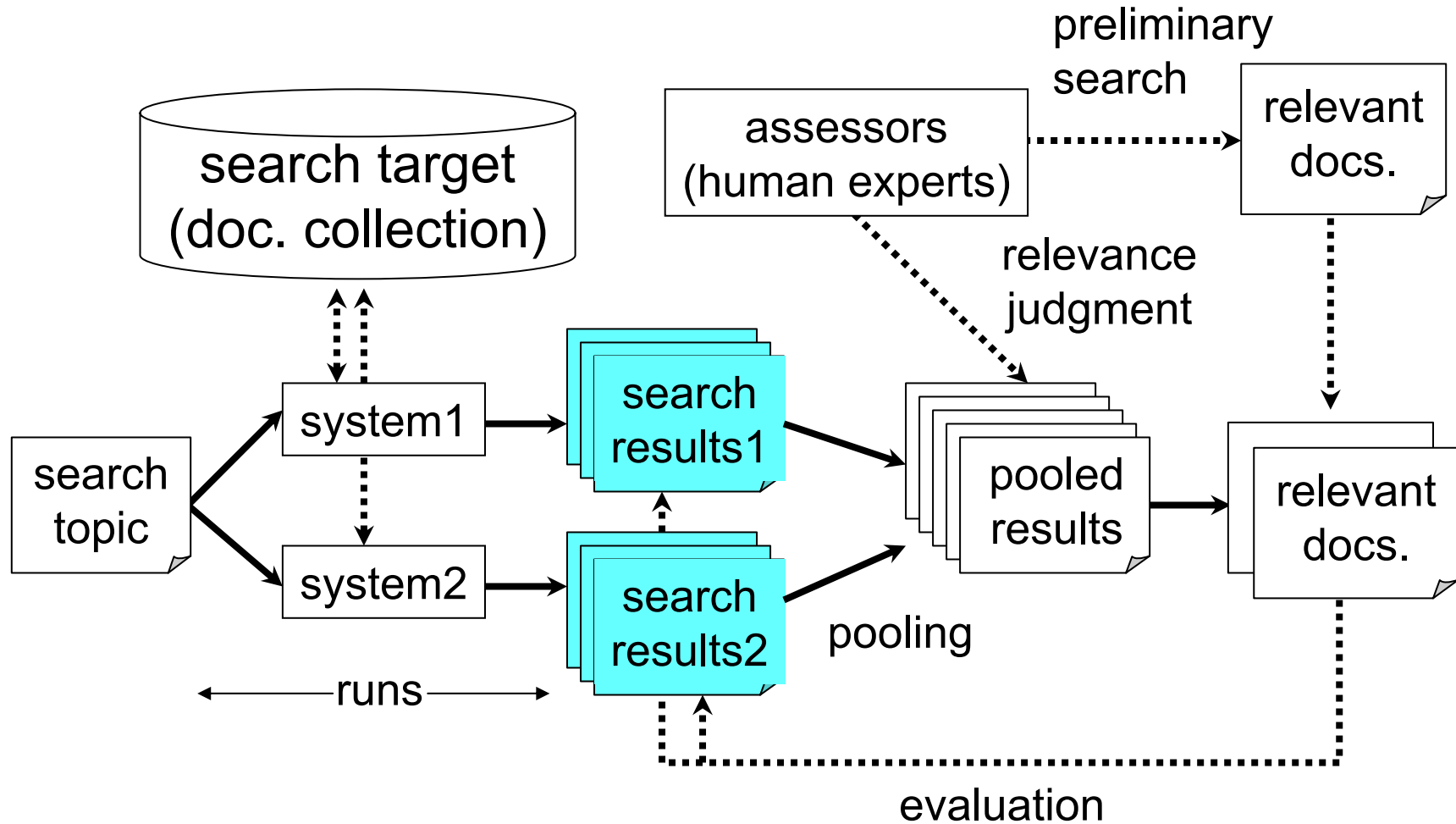
...

</TOPIC>

Relevant documents must be prior art, which had been open to the public **before** the topic patent was filed

Target for invalidation

Process of producing test collection



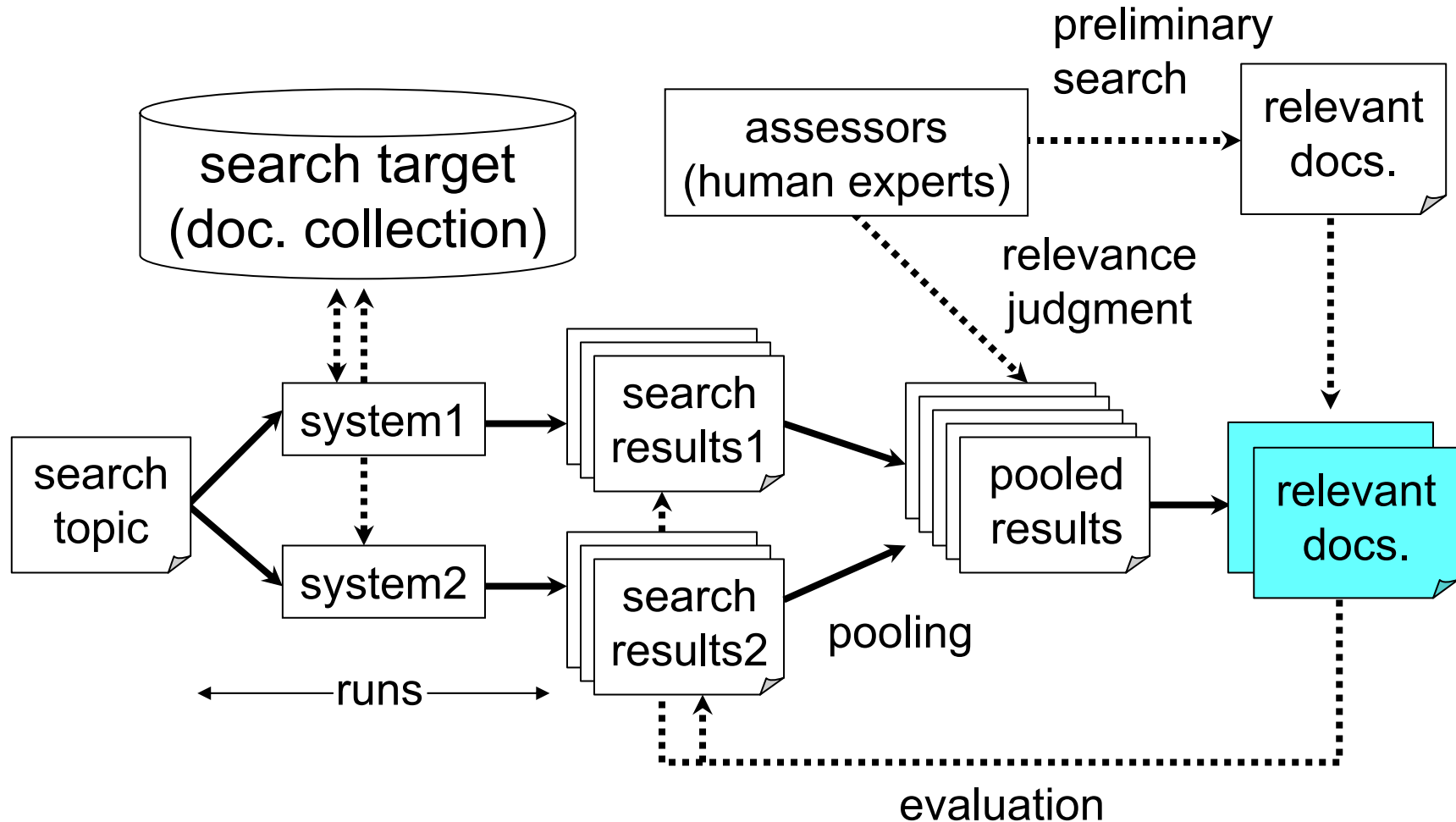
Search results

- For each topic, top 1000 documents are sorted according to the relevance degree
- For each document, passages are also sorted
 - document retrieval and passage retrieval were performed
- Passages are paragraphs determined by applicants
- 110 results were submitted from 8 groups

Example retrieval result

Topic	Passage score	Document ID	Document rank	Document score	System ID
0001	890	1993-123456-5	1	9999	ntc1
0001	870	1993-123456-3	1	9999	ntc1
0001	860	1993-123456-0	1	9999	ntc1
0001	850	1993-123456-12	1	9999	ntc1
0001	990	1995-384359-23	2	9998	ntc1
0001	980	1995-384359-2	2	9998	ntc1
0001	970	1995-384359-8	2	9998	ntc1
0002	890	1994-000002-3	1	9999	ntc1
0002	850	1994-000002-1	1	9999	ntc1
...					

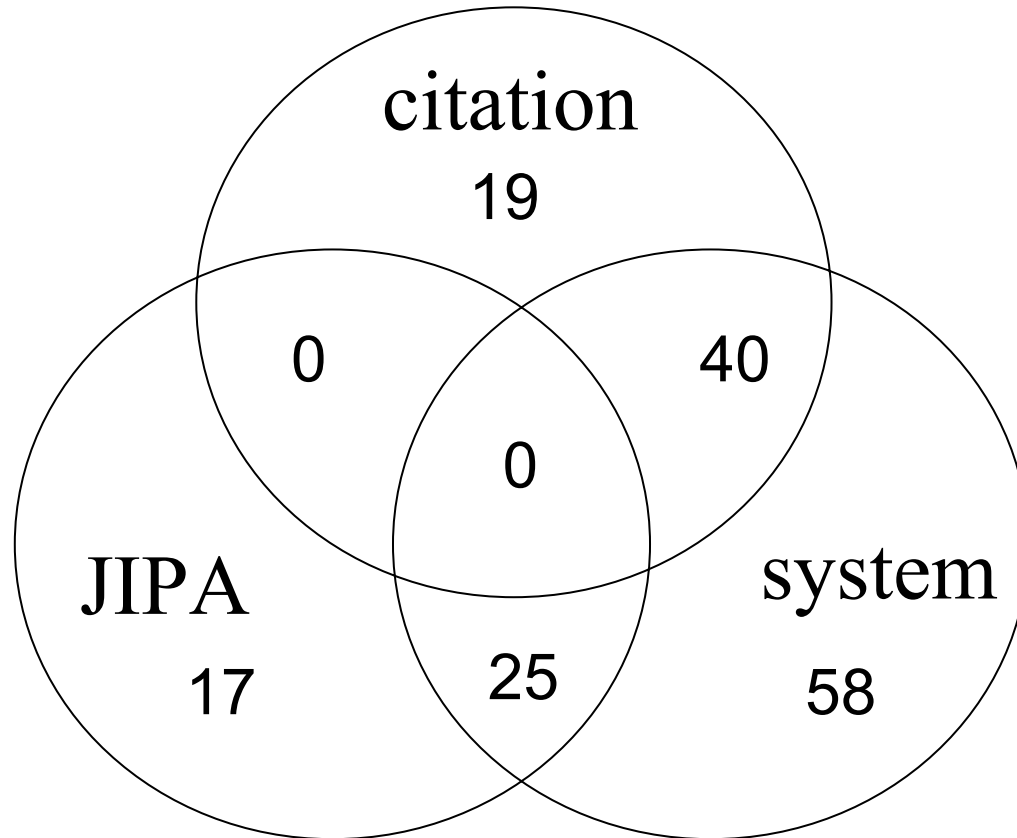
Process of producing test collection



Relevance judgment

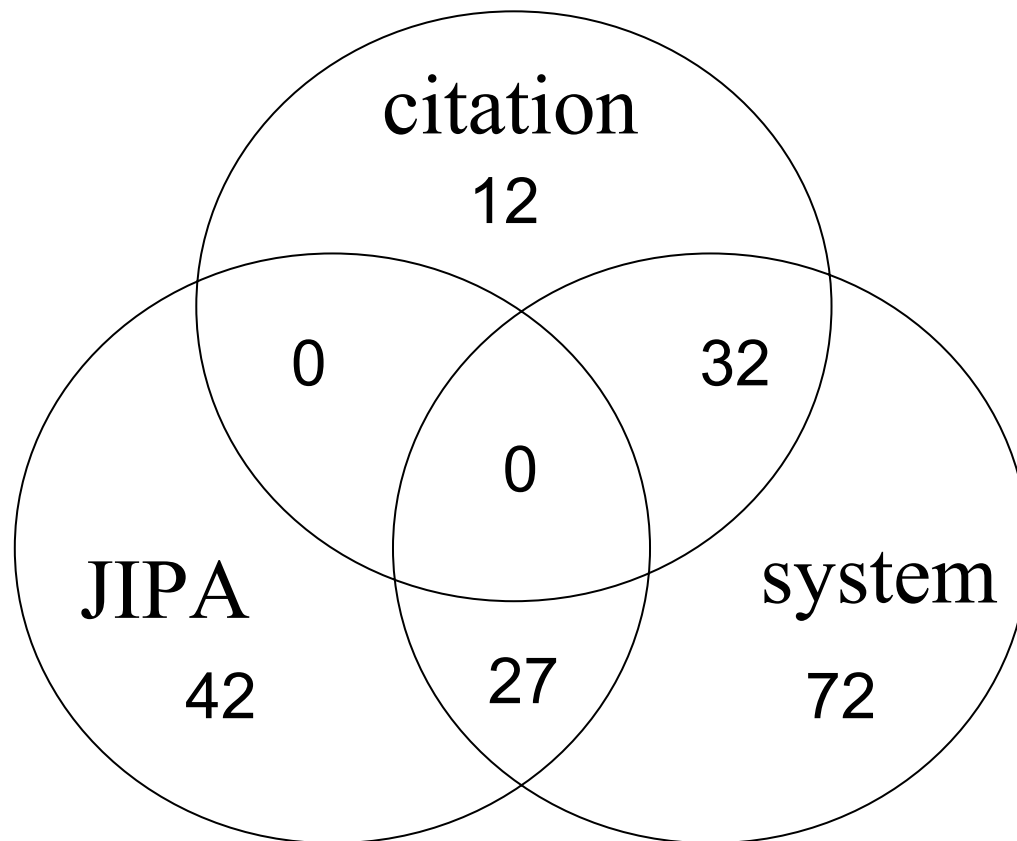
- **Document-based** relevant judgment was performed based on the following two ranks
 - A: patent that can invalidate topic claim
 - B: patent that can invalidate topic claim, when used with other patents
 - (but should be related to most of components)
- Submitted search results were evaluated by mean average precision (MAP)

Details of relevant documents (A)



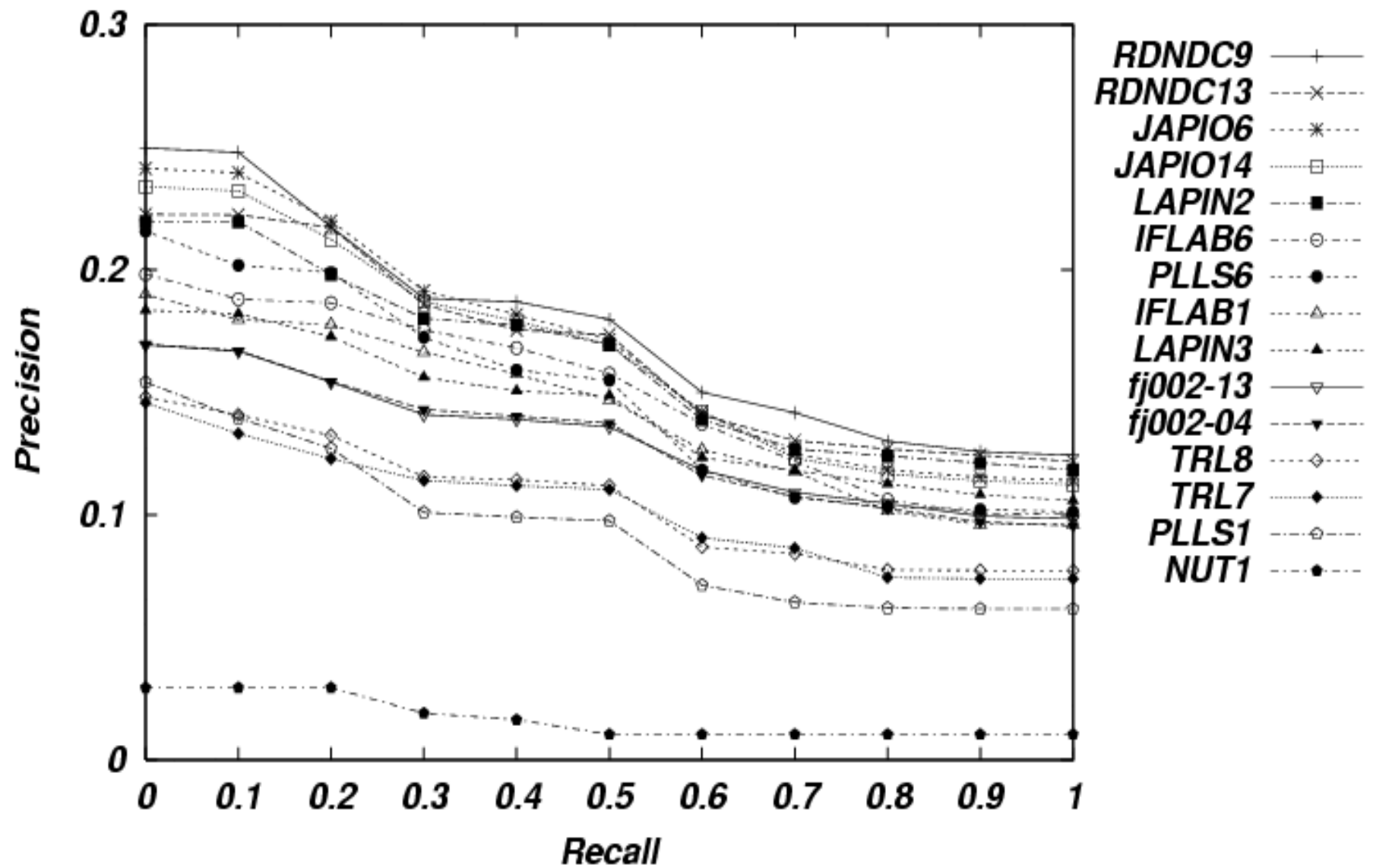
total number of documents is 159

Details of relevant documents (B)



total number of documents is 185

All, Rigid



Formal run results

- no significant difference b/w the results of main topics (34) and additional topics (69)
- please see proceedings for details

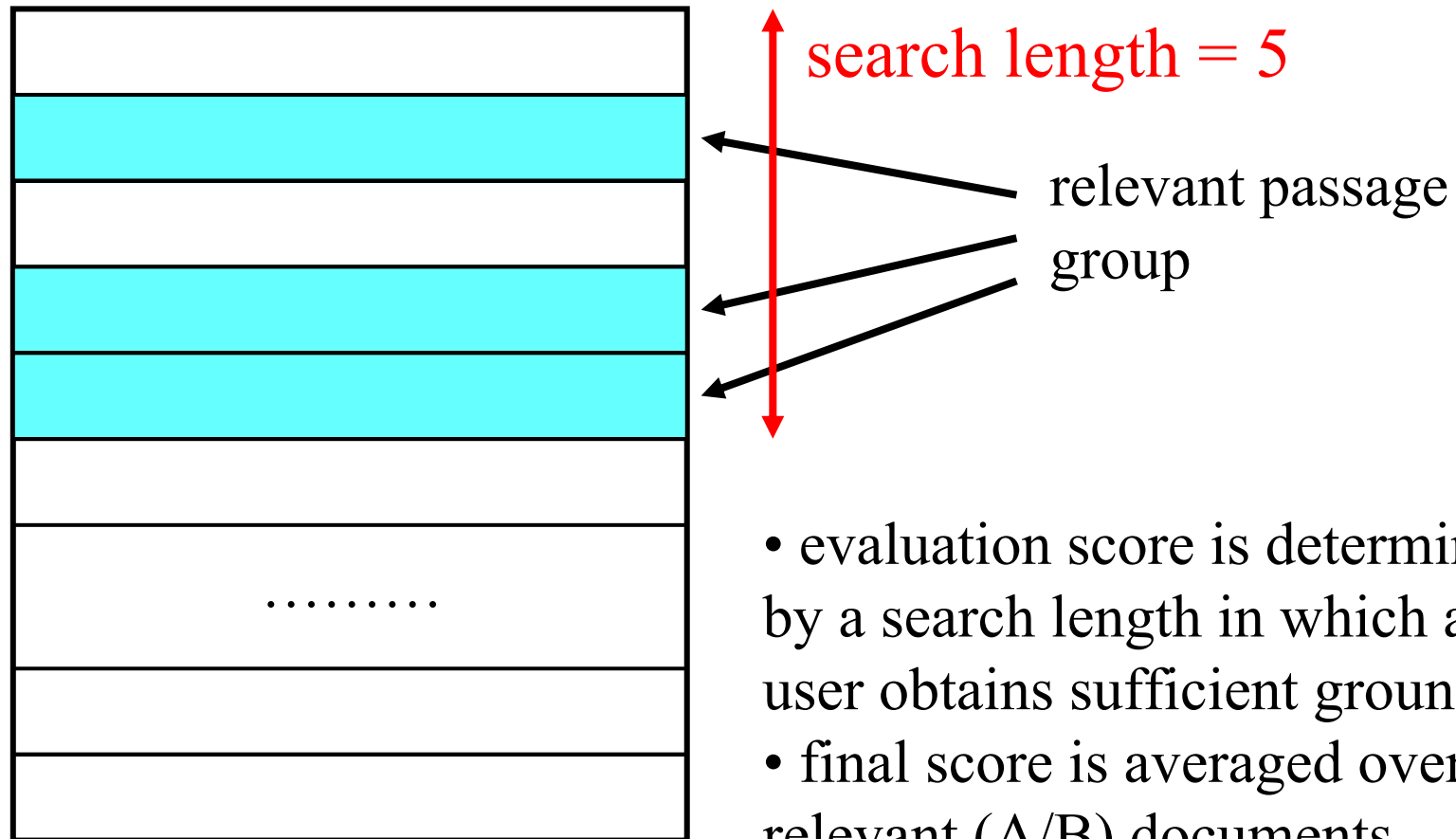
Passage-based relevance judgment

- For each relevant document (either A or B), **passage-based** relevant judgment was performed as follows:
 - if a passage can be grounds to judge the document as relevant, this passage is relevant
 - if a group of passages can be grounds to judge the document as relevant, this **passage group** is relevant
- assessors searched for relevant passages and groups exhaustively

Passage-based evaluation

- Relevant passage group is equally informative as a single relevant passage
- New concept of **combinational relevance** is proposed
- In the conventional evaluation for IR, relevant items (e.g. documents and passages) are independent and therefore combinations are not considered

Example of passage-based evaluation



a relevant document (A or B)

Baseline IR system

- Organizers provided participants with a baseline IR system on the Web
 - return document list in response to a query
 - indented for glass-box comparative evaluation
- Fundamentally, each group was able to participate only by developing front/back-end modules
 - i.e., query processing and passage retrieval
- two groups used the baseline system

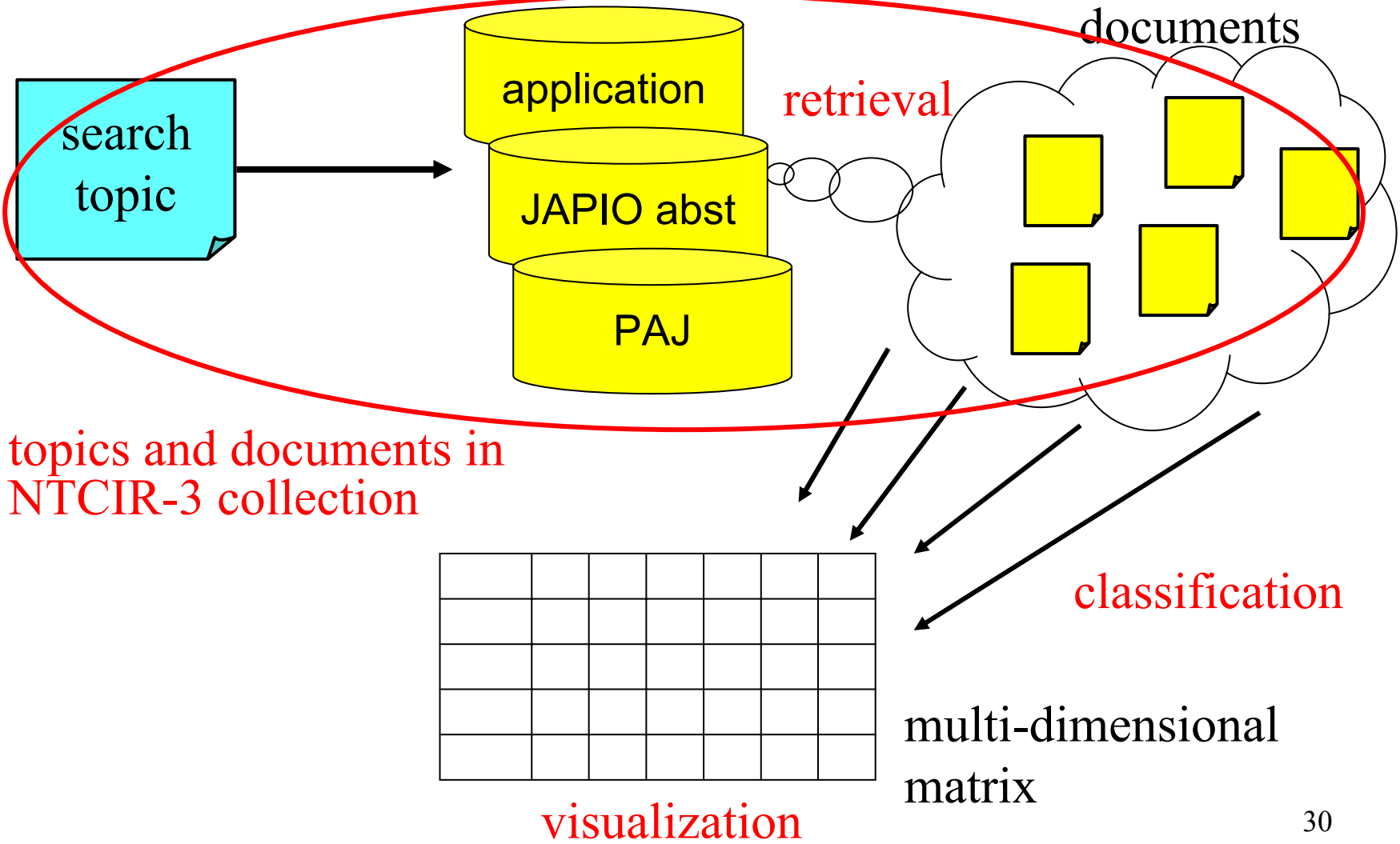
Example methods used by participants

- claim structure analysis
 - dividing claim into subtopics
 - dividing preamble and essential parts
 - different term weights depending on the part
- different usages of classification (IPC)
 - filtering, hierarchy, probabilistic model

NTCIR-4 Workshop (2003-2004)

- NTCIR workshop is in one and half years
 - difficult to explore long-term research topics
- Two different patent tasks were performed
 - invalidity search task: short-term
 - **patent map generation task**: long-term
 - feasibility study (FS) task

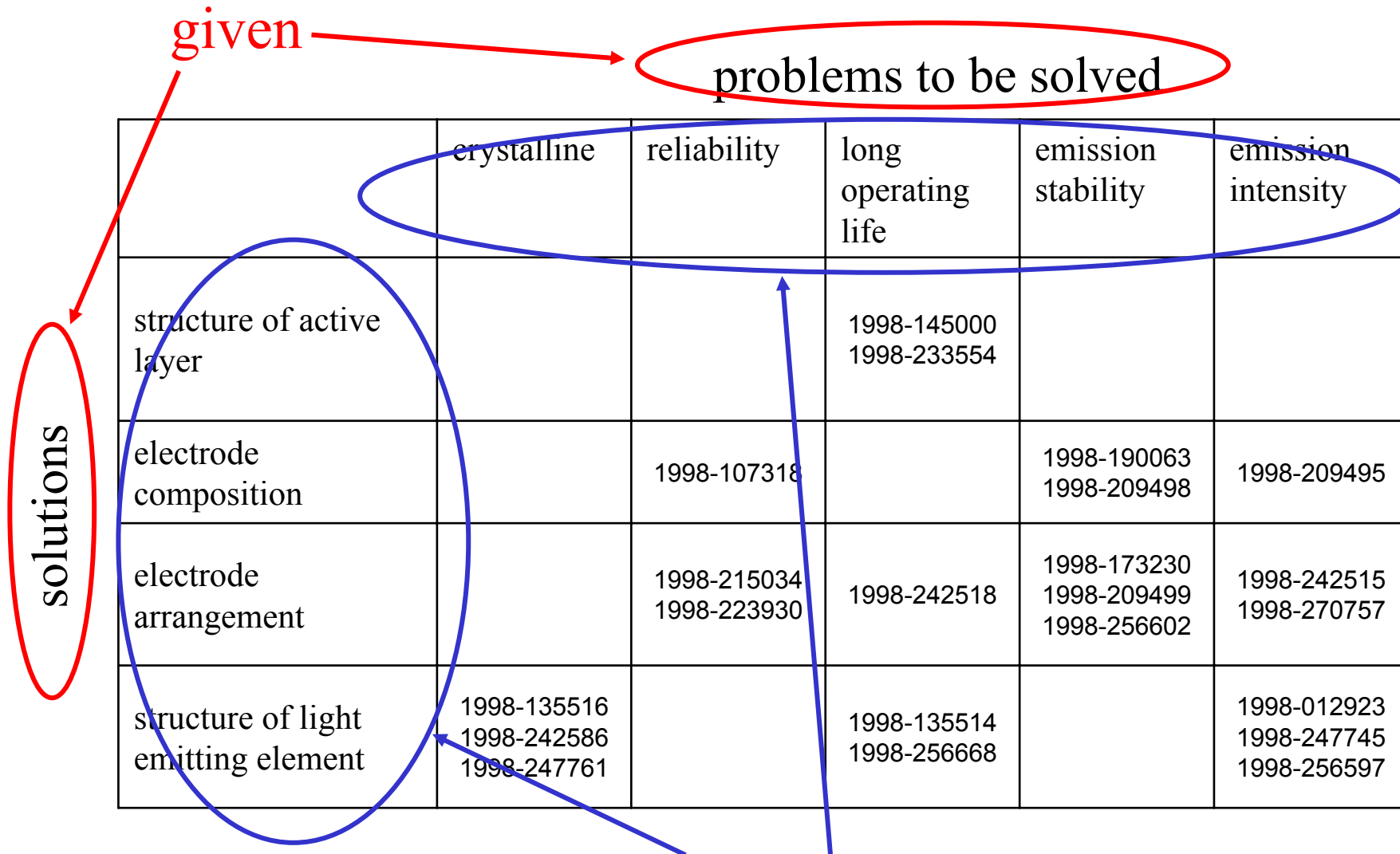
Scenario of patent map generation



Task description

- In principle, given a search topic, relevant patents are retrieved and organized into a multi-dimensional matrix
- In practice, given a search topic, relevant patents and x/y-axes, each participant submits a two-dimensional matrix
 - the number of topics was 6
- human experts evaluated matrix subjectively

Example map (blue light-emitting diode)



participants identify lines and columns

Patent map generation task

- 6 topics were used
 - gasoline direct-injection engine
 - hair care cosmetic products
 - functional carpet
 - blue light-emitting diode
 - solid high-polymer-type fuel cell
 - ultra hydrophilization of plastic surfaces
- human experts produced reference maps and evaluated submitted maps subjectively

Summary

- NTCIR-4 patent collection can be used for
 - retrieval of semi-structured long documents
 - associative patent retrieval
 - passage retrieval
 - classification and text mining (patent map)
- All data will be open to the public after the workshop meeting

Outstanding issues in NTCIR-4

- For invalidity search, the number of relevant documents was inherently small
 - evaluation results can potentially be unreliable
 - to overcome this problem, the number of topics must be increased (cf. question answering task)
- Passage-based evaluation was not used as official result
- The number of participants was small
 - 8 groups (all Japanese groups)

Plan for NTCIR-5

- Two main tasks
- retrieval task
 - using more topics (> 1000)
 - exploring passage-based evaluation
- classification (categorization) task
 - a variation of patent map generation
 - to evaluate machine learning methods
- round-table meeting on June 28, 2004