Information
Retrieval
Facility

TREC-CHEM
Evaluation of Information Retrieval Tools for Chemical Patents and Scientific Articles

Mihai Lupu,
Postdoctoral Research Fellow, IRF

ICIC, October 2009

Outline

- Introduction
- Motivation
- Setup
- Participants
- Methods
- Evaluation
- Lessons
- Your contribution
- Conclusion

TREC Chemistry Information Retrieval

- Organized for the first time this year
- Part of the oldest series of evaluation campaigns
- Aim:
  - Assess the available Chemistry Retrieval tools
  - Generate interest among research groups for this domain
  - Stimulate participation from industry
  - Identify the best measures to be used when evaluating Chemistry Retrieval tools (meta-evaluation)

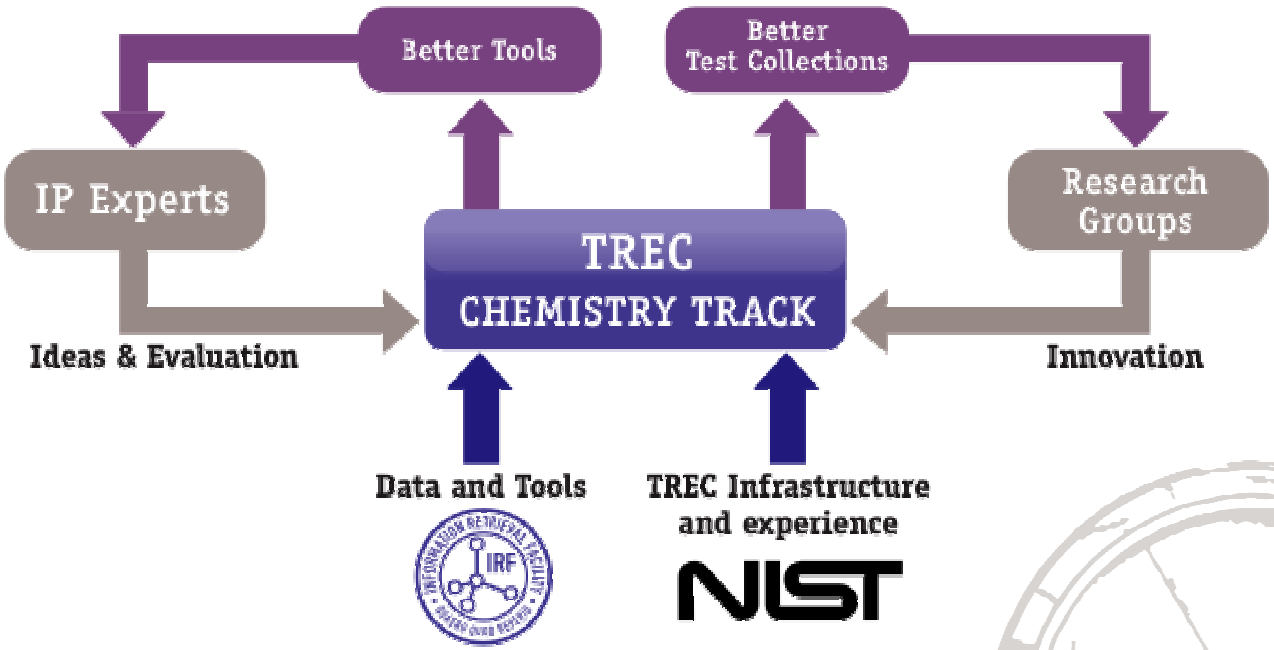
The IRF Mission

- To bridge the gap between the needs of the industry and the academic know-how.
- To bring the latest information retrieval technology to the community of patent professionals and other professional searchers.
- To maintain a facility that enables large scale information retrieval and in depth data processing.

Introduction

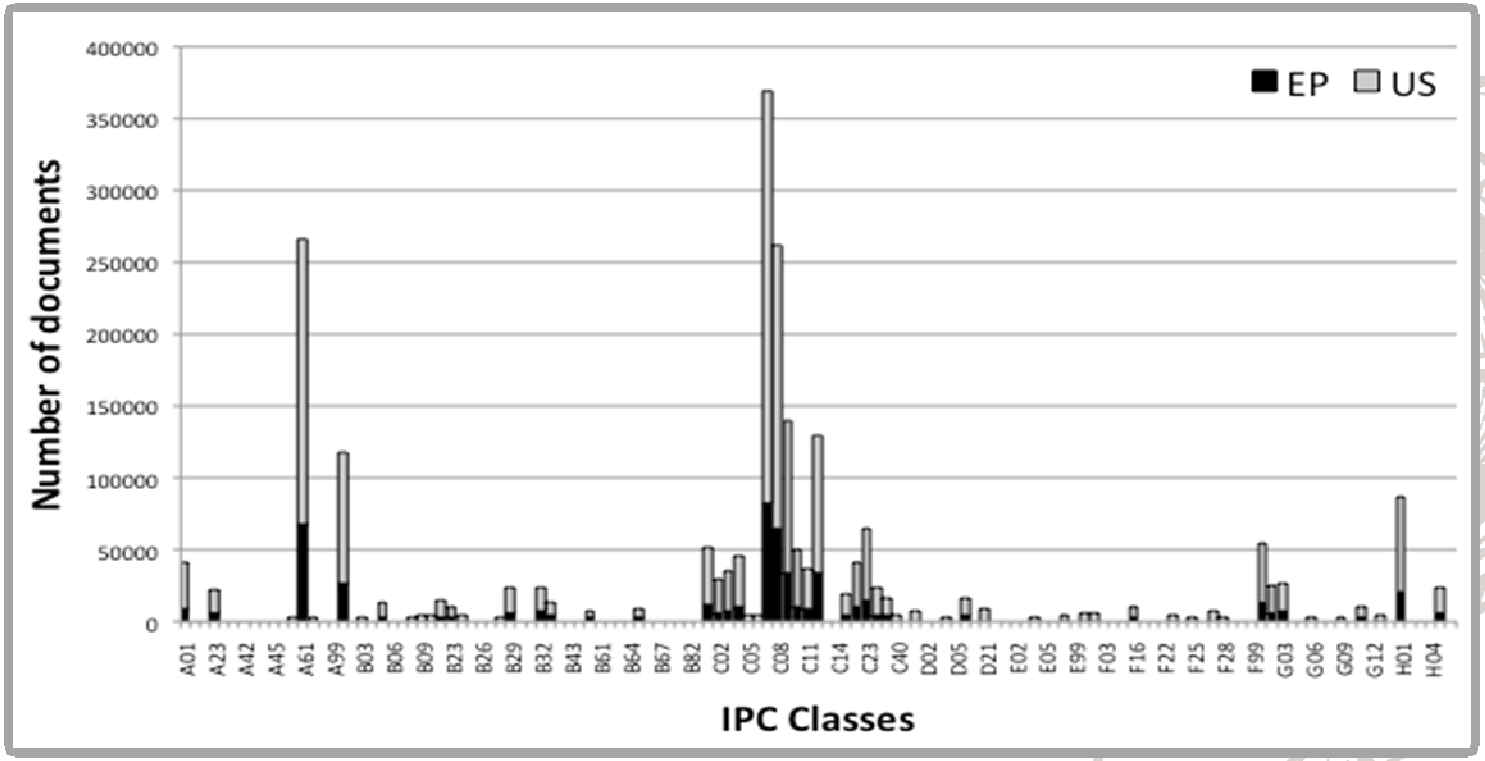
- Collaboration
  - National Institute for Science and Technology (US)
  - University College London (UK)
  - York University (Canada)
- Support from
  - Royal Society of Chemistry
  - Experts in the field
- With the participation of
  - Research groups

Motivation


Better Tools
Better Test Collections
IP Experts
Research Groups
TREC
CHEMISTRY TRACK
Ideas & Evaluation
Innovation
Data and Tools
TREC Infrastructure and experience
NIST

Data

- 1.2 mil. patent files (IRF)
- 59k scientific articles (RSC)
- All English



Tasks

- Technical Survey
  - Search for all potentially relevant documents, in both collections.
  - 18 manually defined and evaluated topics
- Prior Art
  - Search for patents that may invalidate a given patent
  - 1000 automatically created and evaluated topics (1000 patent files)

Participants

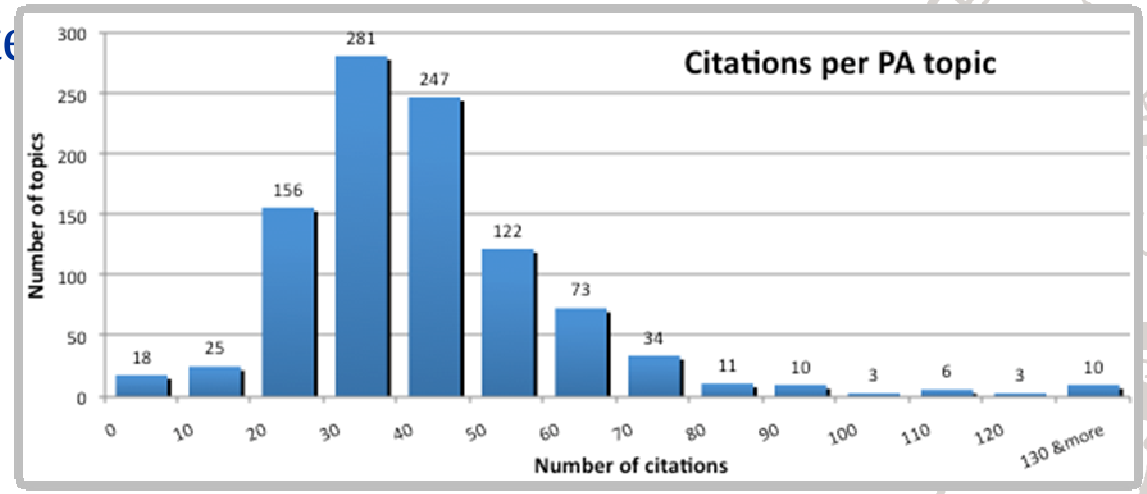
- 15 institutions registered to get the data
  - 6 submitted 31 runs for the TS task:
    - University of Applied Science Geneva, Information Retrieval Laboratory of Dalian University of Technology, Fraunhofer SCAI, Milwaukee School of Engineering, Purdue University, York University
  - 8 submitted 59 runs for the PA topics:
    - University of Applied Science Geneva, **Carnegie Mellon University**, Information Retrieval Laboratory of Dalian University of Technology,**University of Iowa**, Fraunhofer SCAI, Milwaukee School of Engineering, Purdue University, York University

Methods

- Basic vector space model
  - Different sections, weights on each section
  - bm25
- Additional filtering/weighting based on IPC codes
- Linguistic processing
  - Emphasis on NP
- Concept based search
  - Query expansion
  - Using Oscar3, MeSH

Evaluations

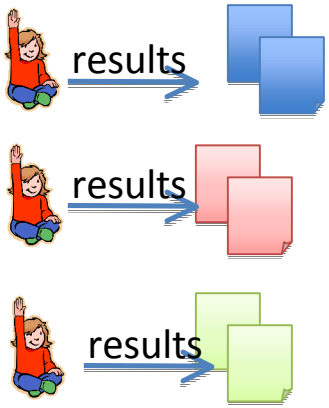
- TS tasks
  - 8 chemistry grad students
  - 5 experts
  - Each topic evaluated by 2 students and 1 expert
- PA tasks
  - Automatically evaluated based on citations within pate

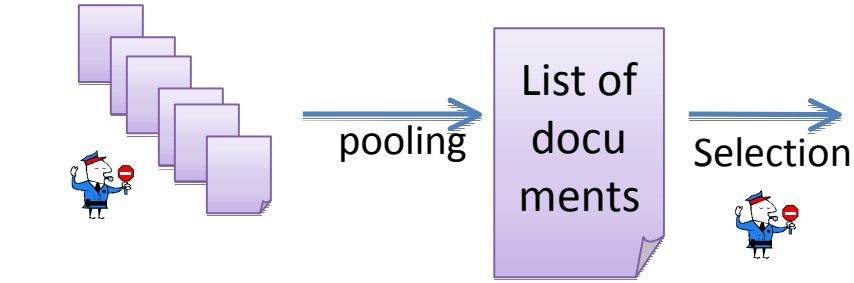


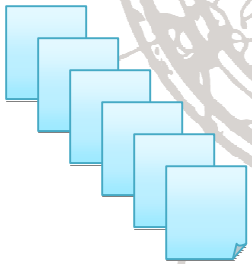
TS Evaluation

Four groups:

1. Organizers
2. Participants
3. Student Evaluators
4. Expert Evaluators

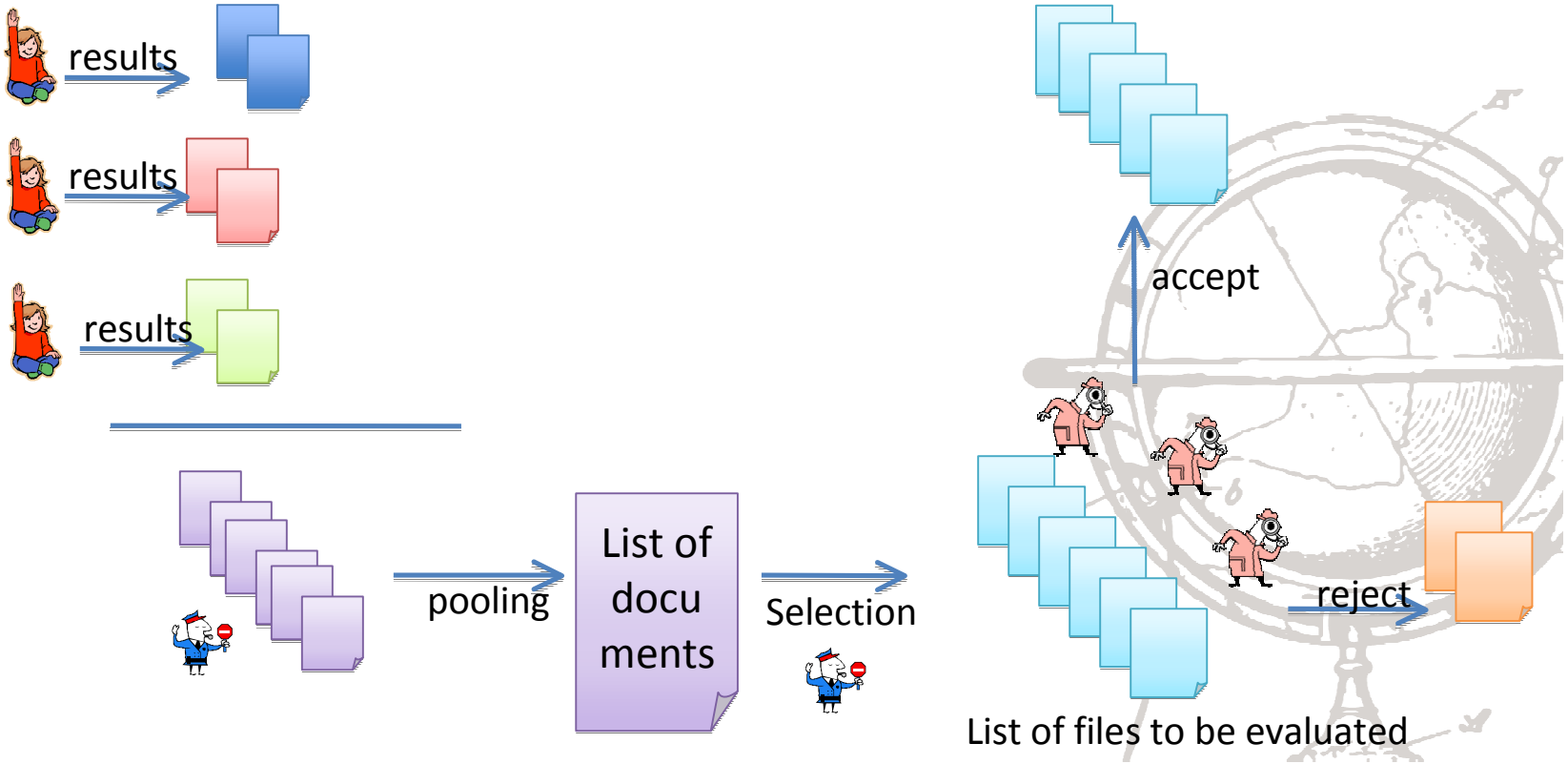




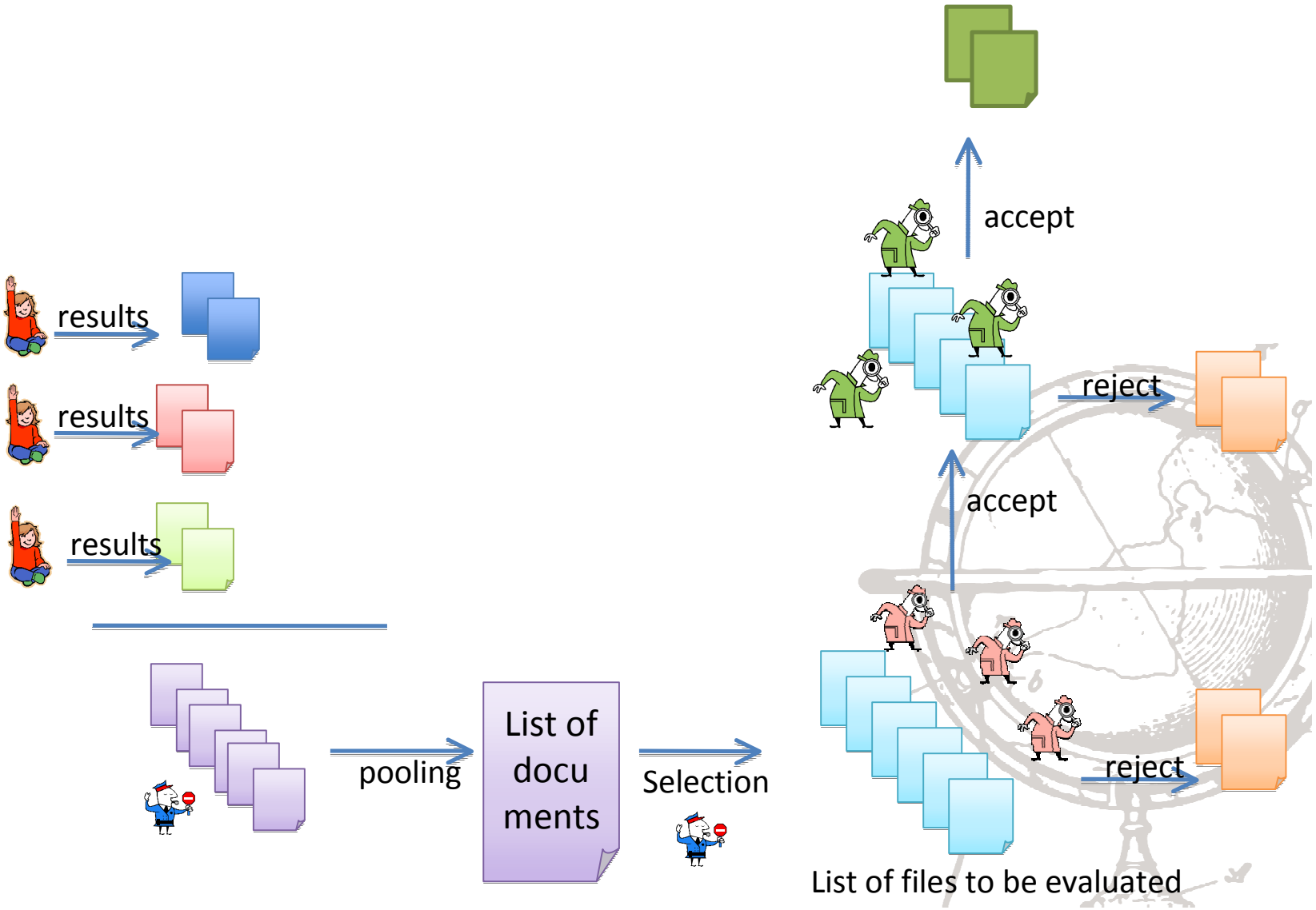


List of files to be evaluated

TS Evaluation



TS Evaluation

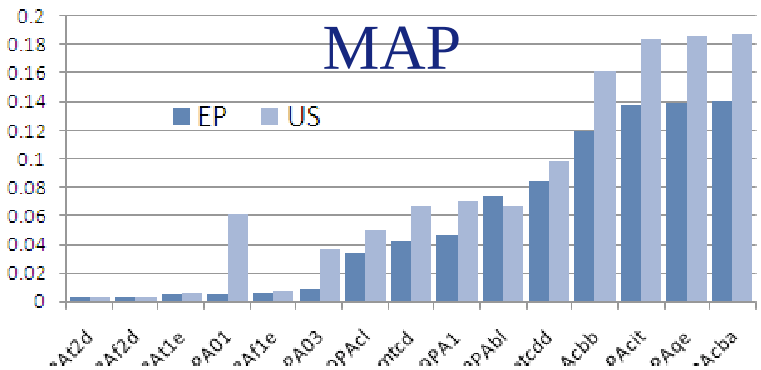


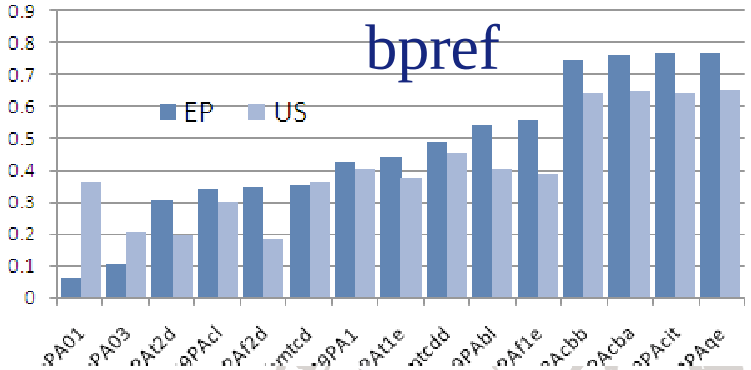
Initial Results

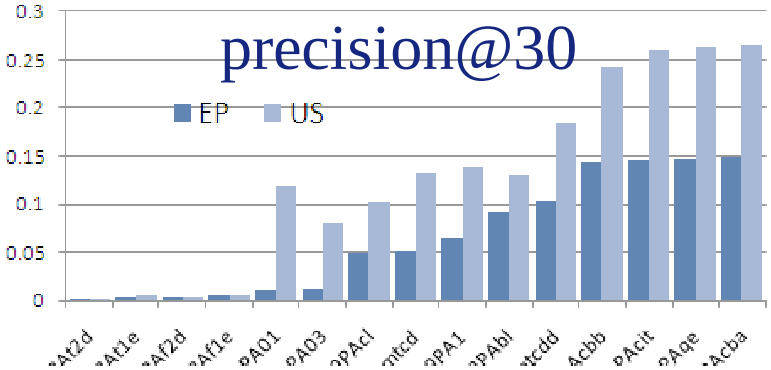
- Manual evaluations have some conflicting results
  - Not more than other manually evaluated topics
- Using entity recognition and synonyms proves successful
  - Some groups manually extended the queries
- “simple methods” seem to also perform well (e.g. Lucene-based, bm25)
  - E.g. for Inferred AP they reach 97% of highest score
    - Disclaimer: results analysis is still ongoing

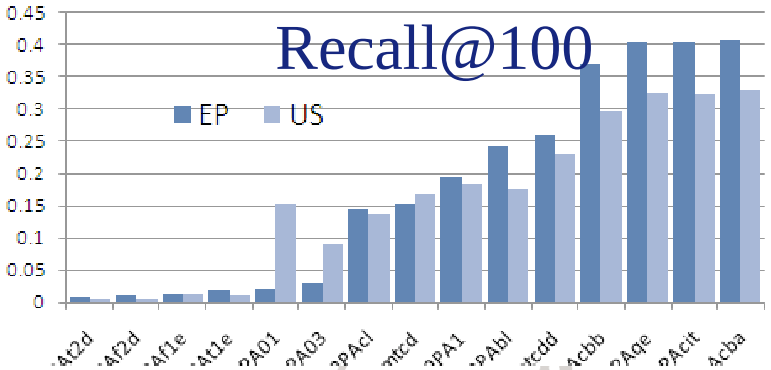
EP – US differences

- Among the 1000 topics in the PA task: 77 from EPO and 923 from USPTO



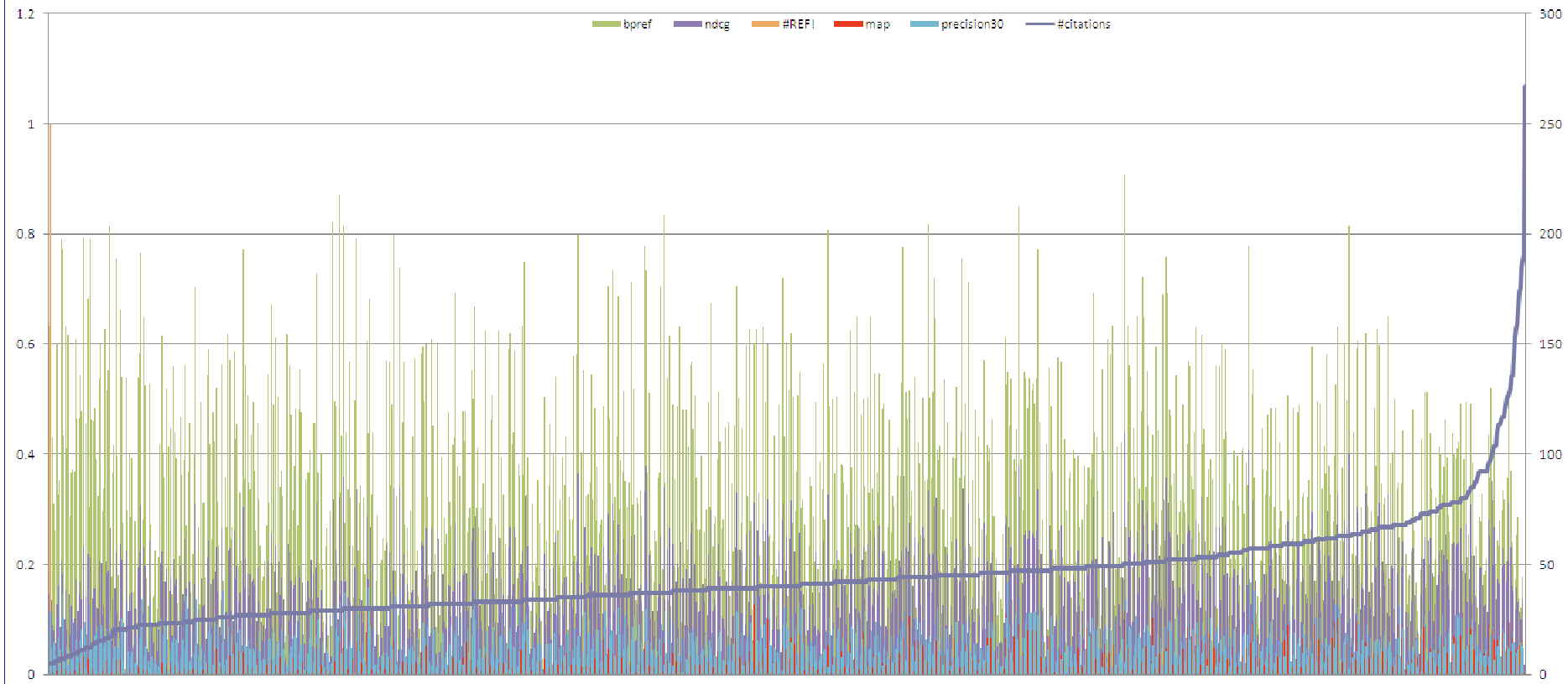






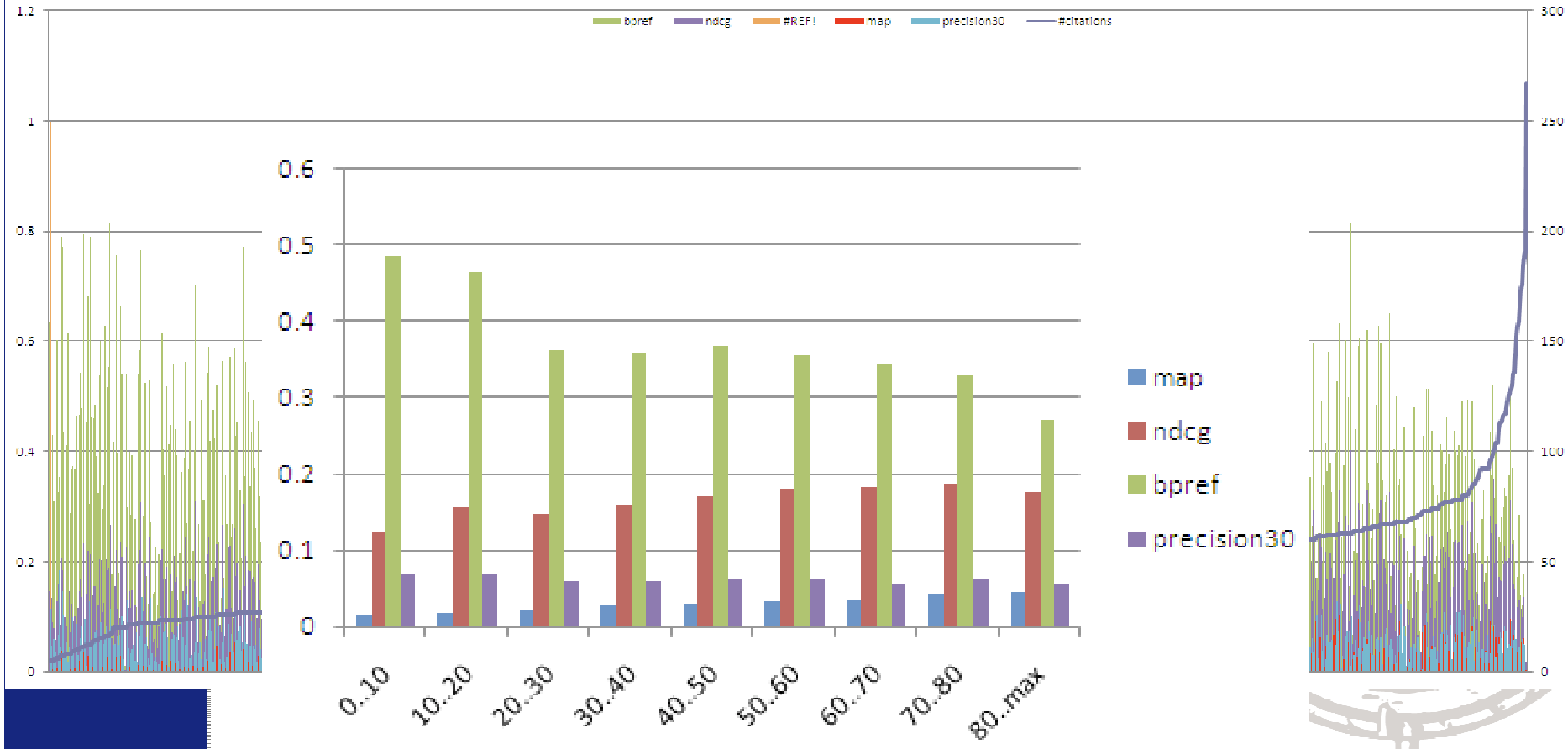
Lessons

- There is bias we need to account for



Lessons

- There is bias we need to account for

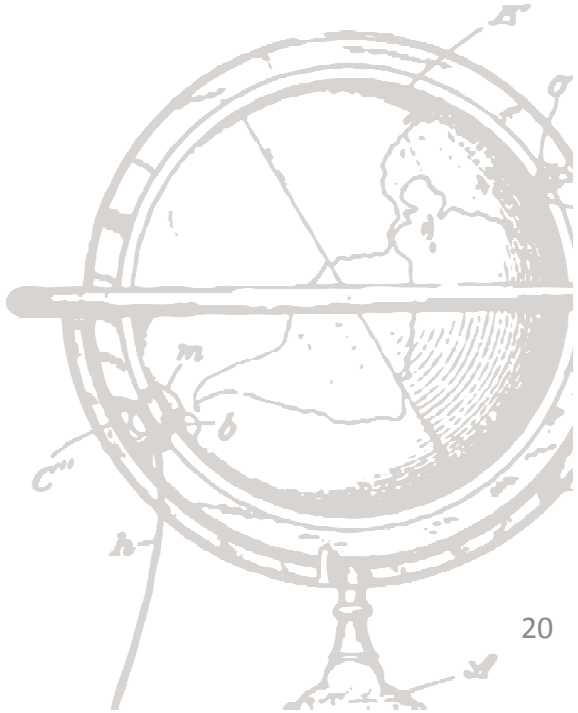


What's next

- Analysis
  - Together with the participants
  - (in Gaithersburg, 17-20th Nov.)
- Decisions for next year
  - Topics
    - Pattern structures
    - Numeric ranges
    - Roles of chemicals
    - Reactions
  - Tasks
    - Image retrieval
    - Entity retrieval
    - Passage retrieval
    - Interactive retrieval

Your contribution

- … is vital
- Whether you are a
  - Patent searcher
  - Data provider
  - Commercial provider



Your contribution

- … is vital… as a patent searcher:
- Proposals of topics to be addressed by participants
  - From your own experience
  - Towards your own goals
- Evaluations of results from the participants
  - After students have filtered out what is clearly not relevant
  - Takes between 1 and 4 hours per topic
    - Up to 6-7h
- Direct involvement with the researchers
  - Interactive task

Your contribution

- … is vital… as a data provider
  - Make [parts of] your data available to researchers
- … is vital… as a commercial tools provider
  - Participate and demonstrate your tools together with the research groups
  - Minimum disclosure necessary (no obligation of releasing sources codes or detailed specification)

Conclusions

- Track report – November 2009
- Final report – February 2010

- Research groups have shown a keen interest
- TREC-CHEM is now a platform ready to be exploited
  - To push research efforts in academia
  - To understand the problems professionals face
  - To have an impartial evaluation of tools

Thank you

- NIST
  - Ellen Voorhees, Ian Soboroff
- RSC
  - Richard Kidd
- Matrixware
  - Rosa Alentorn, Erhard Urban & the team
- Experts
  - Henk Tomas, Monika Hanelt, Tony Trippe and others
- To all of you

Questions please