FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Domain Specific IR
Lecture 5 of 5: Evaluation of DSIR

Mihai Lupu
lupu@ifs.tuwien.ac.at

Russian Summer School on Information Retrieval

August 6-11, 2012

Yaroslavl, Russian Federation

Introduction - IR Evaluation

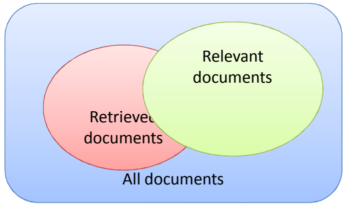
- *"Efficient and effective system"*
- Time and space: efficiency
  - Generally constrained by pre-development specification
    - E.g. real-time answers vs. batch jobs
    - E.g. index-size constraints
  - Easy to measure
- Good results: effectiveness
  - Harder to define --> more research into it
- And…

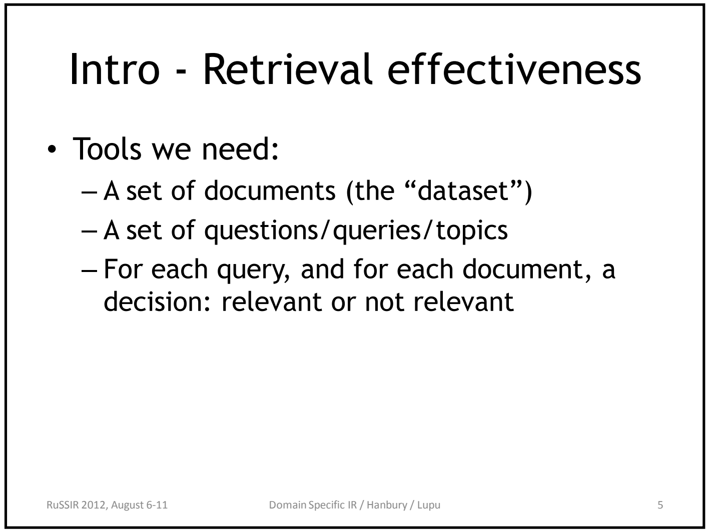
- User studies
  - Does a 2% increase in some retrieval performance measure actually make a user happier?
  - Does displaying a text snippet improve usability even if the underlying method is 10% weaker than some other method?

  - Hard to do
  - Mostly anecdotal examples
  - IR people don't like to do it (though it's starting to change)

Intro - Retrieval effectiveness

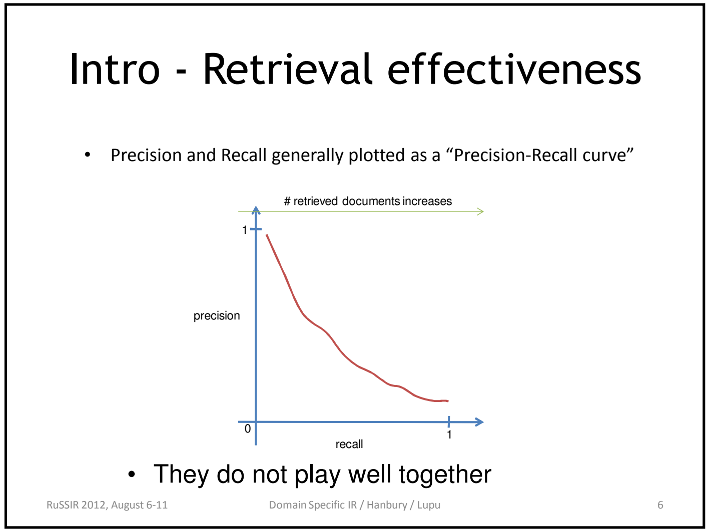
- Precision
  - rel ret /ret
- Recall
  - rel ret/rel





Intro - Retrieval effectiveness

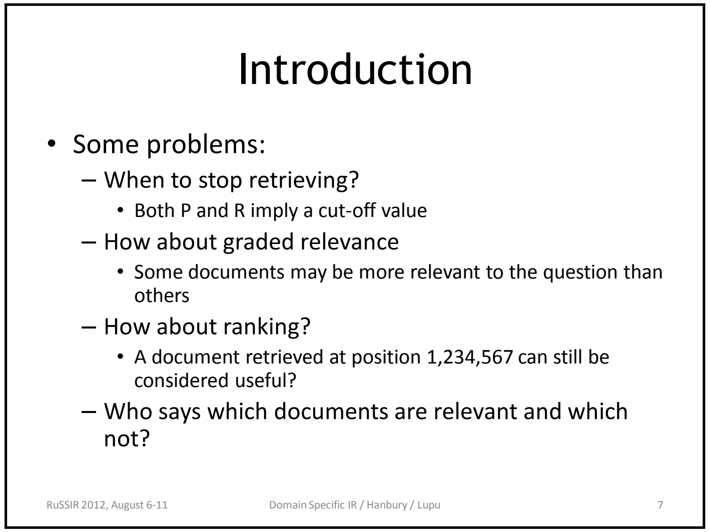
- Tools we need:
  - A set of documents (the "dataset")
  - A set of questions/queries/topics
  - For each query, and for each document, a decision: relevant or not relevant



Intro - Retrieval effectiveness

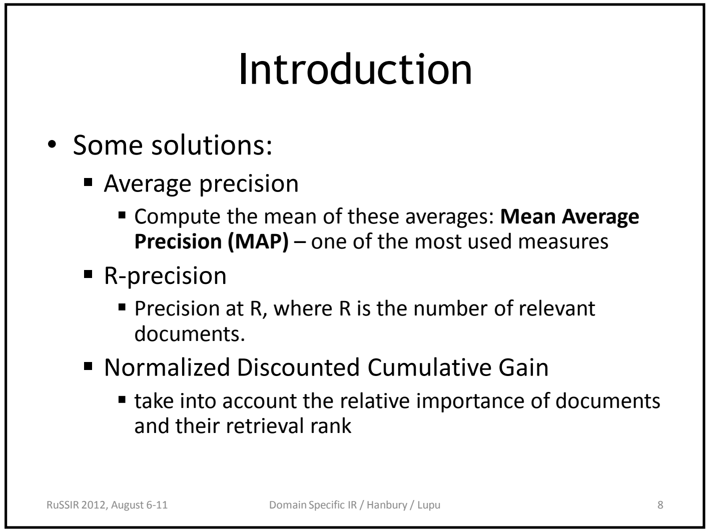
- Precision and Recall generally plotted as a "Precision-Recall curve"

- They do not play well together



Introduction

- Some problems:
  - When to stop retrieving?
    - Both P and R imply a cut-off value
  - How about graded relevance
    - Some documents may be more relevant to the question than others
  - How about ranking?
    - A document retrieved at position 1,234,567 can still be considered useful?
  - Who says which documents are relevant and which not?



Introduction

- Some solutions:
  - Average precision
    - Compute the mean of these averages: **Mean Average Precision (MAP)** – one of the most used measures
  - R-precision
    - Precision at R, where R is the number of relevant documents.
  - Normalized Discounted Cumulative Gain
    - take into account the relative importance of documents and their retrieval rank

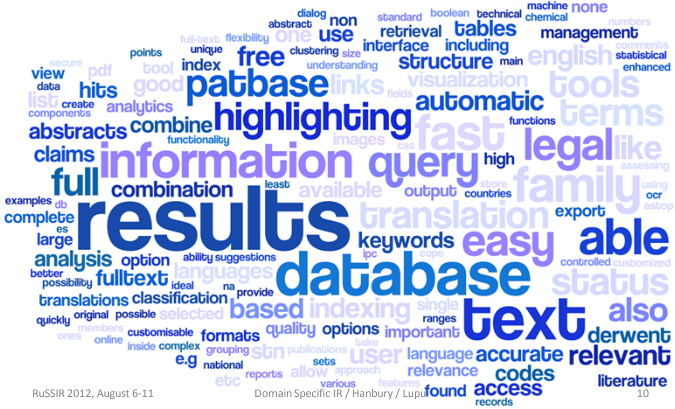
Introduction

- Some more measures
  - There are tens of IR measures!
  - trec_eval is a little program that computes many of them
    - 37 in v9.0, many of which are multi-point (e.g. Precision @10, @20…)
  - http://trec.nist.gov/trec_eval/
  - "there is a measure to make anyone a winner"
    - Not really true, but still…

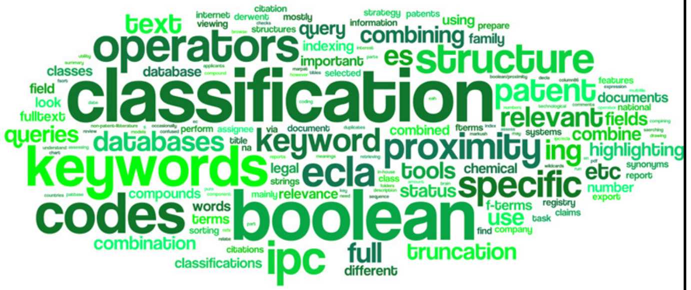
The use/need of images

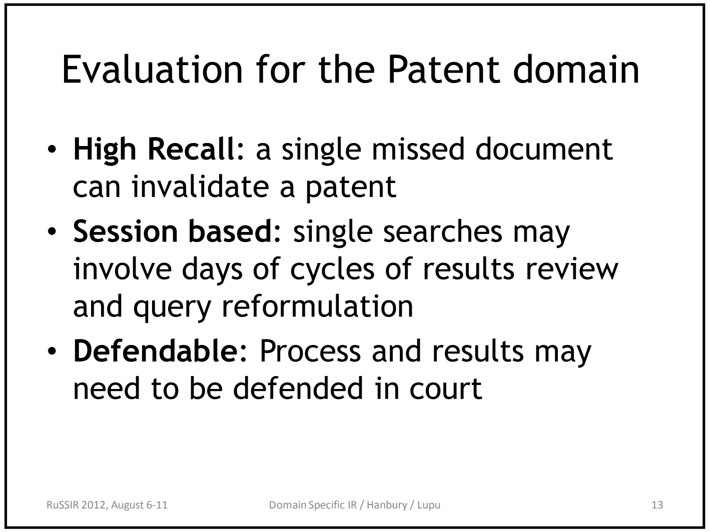
Question 1	Question 2	Question 3 - rate between 1(Not useful)-5(very useful)					Question 4
In your experience, which subclasses of the IPC require image search? Give at most 3 and for each provide an example. Obs. Exclude chemical structure recognition	For each of the IPC subclasses listed above, what are some criteria for two images to be considered similar?	Find text in figures, recognize and index it	Link numbers found in images with those in text	Recognize the view perspective difference of two images	Recognize the different types of images	Image enhancement (part highlighting)	Do existing tools provide detailed information about the search report?
Mechanical things of all kind , electrical circuitry. Other fields also.	must recognize components (could also think of them as 'concepts') and understand how they are related to each other	4.29	5.00	3.43	3.83	4.67	Partially. very few do, some are internal
			The numbers in different figures are generally not unique	This is generally considered unrealistic or feasible only in conditions in which humans are vastly more capable	some utility in very specific cases	very useful, but also considered difficult due to partially obscured componets	

Could you describe your ideal patent search system?



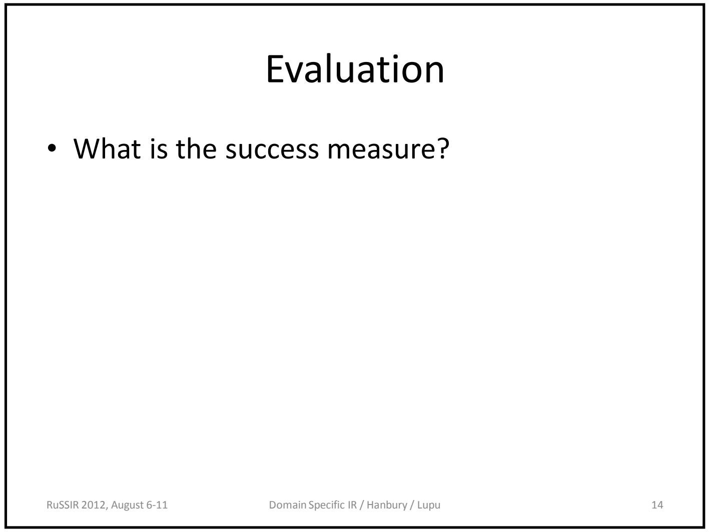
What are the main search features you use to complete a search task?





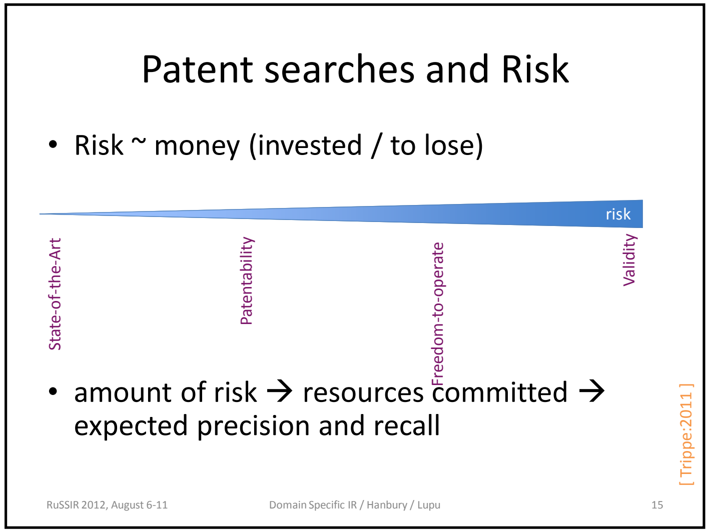
Evaluation for the Patent domain

- **High Recall:** a single missed document can invalidate a patent
- **Session based:** single searches may involve days of cycles of results review and query reformulation
- **Defendable:** Process and results may need to be defended in court



Evaluation

- What is the success measure?

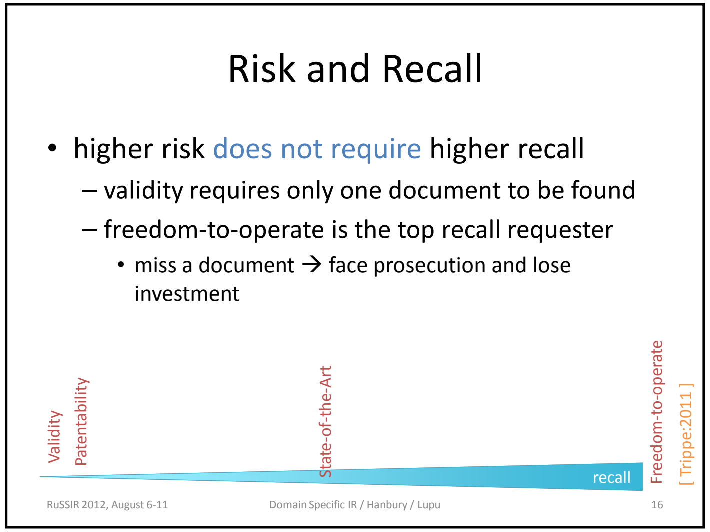


Patent searches and Risk

- Risk ~ money (invested / to lose)

- amount of risk → resources committed → expected precision and recall

[Tripe:2011]



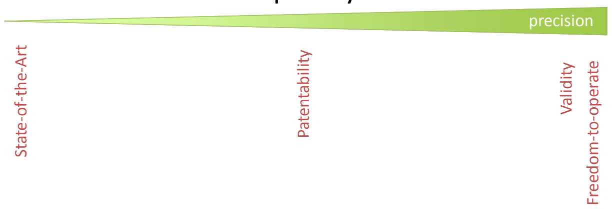
Risk and Recall

- higher risk does not require higher recall
  - validity requires only one document to be found
  - freedom-to-operate is the top recall requester
    - miss a document → face prosecution and lose investment

[Tripe:2011]

Risk and Precision

- match almost completely



[Trippe:2011]

Practice in the IP world

- Commercial world
  - no extensive experimentation
  - based on practice and experience
  - highly competitive
    - and yet often collaborative
      - not one tool is ever declared the best
  - Source of articles
    - *World Patent Information* Journal

Example of evaluation

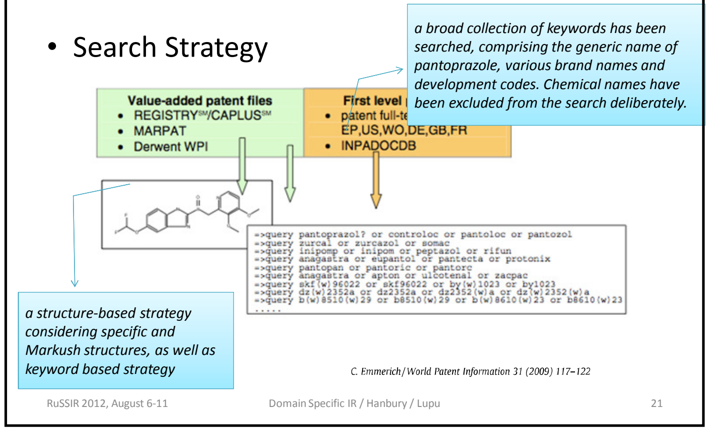
- [Emmerich:2009]
  - case study analysis
    - pharma
  - compares
    - first-level patent data
    - value-added patent information
      - Chemical Abstracts and Thomson Scientific
- background:
  - valued-added patent information sources are the incumbents here

Case study 1

- a prior art search for pantoprazole focusing on worldwide literature
  - particular interest:
    - product protection
    - manufacturing processes
    - formulation/combination patents
    - methods of disease treatment

Case study 1

- Search Strategy



a broad collection of keywords has been searched, comprising the generic name of pantoprazole, various brand names and development codes. Chemical names have been excluded from the search deliberately.

```
=>query pantoprazol? or controloc or pantoloc or pantozol
=>query zurcal or zurcazol or somac
=>query inipomp or inipom or peptazol or rifun
=>query anagastra or eupantol or pantecta or protonix
=>query pantopan or pantoric or pantorc
=>query anagastra or apton or ulcotenal or zacpac
=>query skf(w)96022 or skf96022 or by(w)1023 or by1023
=>query dz(w)2352a or dz2352a or dz2352(w)a or dz(w)2352(w)a
=>query b(w)8510(w)29 or b8510(w)29 or b(w)8610(w)23 or b8610(w)23
.....
```

a structure-based strategy considering specific and Markush structures, as well as keyword based strategy

C. Emmerich / World Patent Information 31 (2009) 117–122

Case Study 1

- Results
  - Value-added search: 587 inventions
    - of these
      - each source had at least 3.6% unique results (one had 19.6%)
      - overlaps: 68.8%
  - Full-text search : 1097 inventions
    - not found: 117 inventions
    - new : 651 inventions

Case study 1

- Analysis
  - comparison of precision
    - Value-added data: <1% false hits
    - Full-text search: >30% non relevant
  - why different results:
    - value-added data:
      - procedural differences in indexing (not everything is indexed: not all patent documents and not all chemical formulas)
      - coverage
    - full-text search:
      - coverage
    - value-added data vs full-text search
      - + Asian patent publications with no English text
      - compositions with could be used to deliver this and other drugs
      - decision to index only some structures

Case study 1

- Analysis
  - failures of full-text
    - key patents cannot be found due to
      - representation as a chemical structure only (potentially part of a Markush structure)
    - not standardized chemical names
    - misspellings

Case Study 1

- Conclusions of this case study
  - multiple sources need to be used
  - a set of characteristics of each tool/source
- Our conclusions based on this study
  - 1 query
  - impossible to repeat (not enough details)
  - evaluation merges collection coverage and query capabilities
  - –
  - –
  - –

Case Study 2

- [Annies:2009] reviews
  - search and display capabilities of e-journal search sites
  - value for full-text prior art analysis
- Target data/systems
  - e-journals' publishers' websites
  - ! many discarded from the beginning
    - *"many search sites are not useful for professional prior art searching, since they lack advanced search and display features critical for fast and effective retrieval and evaluation"*

Case Study 2

- minimum set of requirements
  1. site should cover a larger number of e-journals
  2. provide advanced search options (e.g. at least Boolean logic with wildcards)
  3. provide advanced display features (e.g. at least search keywords highlighting)
- out of 200 sites available to the author, 4 fulfilled these 3 basic requirements

Case Study 2

- search features analysis
  - how query can be defined
  - search by fields?
  - other features: date filtering, phrase searching, stemming, wildcards, citation search, proximity operators
- display features analysis
  - keyword highlighting on different colors based on concepts
- other features
  - save/history options
  - RSS feeds and search alerts
  - open access
- chemical structure search
  - none of the 4
  - 2 of the 200

Case Study 2

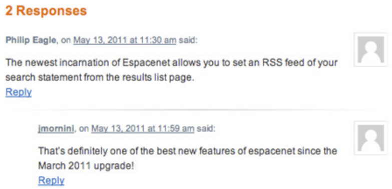
- Conclusions of this case study
  - e-journal search options offered by publishers are insufficient for professional search
    - why?
      - patent information professionals search for rather hidden information
      - they apply more complex search strategies for comprehensive retrieval
  - following aspects found problematic:
    - search and display features limited
    - spread of journals across non-cooperating publishers

Case Study 2

- Our conclusions on this case study
  - absolutely no mentioning of search effectiveness
  - starting point is a predefined wish list
  - 'evaluation' is all-encompassing (from coverage, to search, to display)
  - –
  - –

Other evaluations

- Community based
  - e.g.Intellogist, PIUG wiki, LinkedIn Groups
  - evaluation is done 'in the wild'
  - experiences shared
- e.g.



2 Responses

Philip Eagle, on May 13, 2011 at 11:30 am said:

The newest incarnation of Espacenet allows you to set an RSS feed of your search statement from the results list page.

Reply

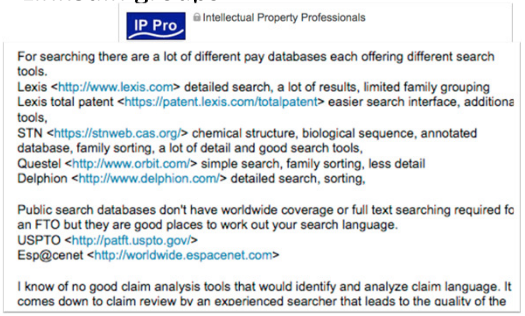
jmornini, on May 13, 2011 at 11:59 am said:

That's definitely one of the best new features of espacenet since the March 2011 upgrade!

Reply

Other evaluations

- LinkedIn groups



IP Pro Intellectual Property Professionals

For searching there are a lot of different pay databases each offering different search tools.
Lexis <http://www.lexis.com> detailed search, a lot of results, limited family grouping
Lexis total patent <https://patent.lexis.com/totalpatent> easier search interface, additiona
tools,
STN <https://stnweb.cas.org/> chemical structure, biological sequence, annotated database, family sorting, a lot of detail and good search tools,
Questel <http://www.orbit.com/> simple search, family sorting, less detail
Delphion <http://www.delphion.com/> detailed search, sorting,

Public search databases don't have worldwide coverage or full text searching required fo
an FTO but they are good places to work out your search language.
USPTO <http://patft.uspto.gov/>
Esp@cenet <http://worldwide.espacenet.com>

I know of no good claim analysis tools that would identify and analyze claim language. It comes down to claim review bv an experienced searcher that leads to the qualitv of the

Outline

- Practice in the IP world
- Practice in the IR world
  - Useful research
    - evaluating relevance feedback
    - evaluating interaction

Practice in the IR world

A World of Difference

- it looks at:
  1. effectiveness of the core engine
  2. repeatability of experiments
  3. statistical significance of experiments

  […]

  20. user interface

Practice in the IR World

- organize large evaluation campaigns
  - TREC
  - CLEF
  - NTCIR
  - INEX
  - FIRE
  - …

TREC - Topics

- For TREC, topics generally have a specific format (not always though)
  - <ID>
  - <title>
    - Very short
  - <description>
    - A brief statement of what would be a relevant document
  - <narrative>
    - A long description, meant also for the evaluator to understand how to judge the topic

TREC - Topics

- Example:
  - <ID>
    - 312
  - <title>
    - Hydroponics
  - <description>
    - Document will discuss the science of growing plants in water or some substance other than soil
  - <narrative>
    - A relevant document will contain specific information on the necessary nutrients, experiments, types of substrates, and/or any other pertinent facts related to the science of hydroponics. Related information includes, but is not limited to, the history of hydro- …

CLEF

- Cross Language Evaluation Forum
  - From 2010: Conference on Multilingual and Multimodal Information Access Evaluation
  - Supported by the PROMISE Network of Excellence
- Started in 2000
- Grand challenge:
  - Fully multilingual, multimodal IR systems
    - Capable of processing a query in any medium and any language
    - Finding relevant information from a multilingual multimedia collection
    - And presenting it in the style most likely to be useful for the user

NTCIR

- Started in 1997, but organized every 1.5 years
- The first to look at Patent data (in 2001/2002)
- Other tracks:
  - Japanese / Cross-language retrieval
  - Web Retrieval
  - Term extraction
  - QA
    - Information Access Dialog
  - Text summarisation
  - Trend information
  - Opinion analysis

Evaluation campaigns

- two types of ‘interesting’ campaigns
  - those which use patent data and simulate patent search
  - those which evaluate IR features identified as useful by patent professionals
    - e.g.
      - session-based search
      - relevance feedback

CLEF-IP

- since 2009
  - to encourage and facilitate research in the area of patent retrieval
  - to create a large test collection
- focus
  - ad hoc search (Prior Art task)
  - from 2010: classification
- the collection
  - EPO documents (English, French and German)
  - increasing every year (start from 1.9mil in 2009)
  - added WIPO documents in 2011

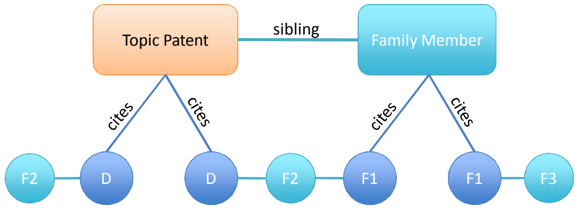
CLEF-IP

- Topics
  - patent documents
    - 2009: topic = a mixture of all documents pertaining to a patent - wrong
    - form 2010: topic = an application document – better
  - selection process
    - defined topic pool (recent documents)
    - textual content must be present in the publication
    - the topic patent must have at least 3 citations in their search reports

CLEF-IP

- Relevance assessments
  - different degrees of relevance
    - from Applicant – less important
    - from Search Report (examiner) – important
    - from opposition procedure (competitor)- most important
  - by definition incomplete

CLEF-IP Relevance judgments



CLEF-IP

- Evaluation procedure and measures
  - pretty much the same as all other IR evaluation campaigns
  - one new measure introduced in 2010
    - PRES [Magdy:2010]
      - recall oriented: lenient on runs which return lots of relevant documents but not necessarily highly ranked, hard to systems which return only a few relevant documents at the top

CLEF-IP 2012

- Claims to Passage
- A topic is now a set of claims, exactly as mentioned in the search report
- The 'gold standard' is now exactly what the examiner indicated
- The evaluation is done both at document (PRES) and at passage level (MAP).
- Come to CLEF 2012 (Rome, Sept 2012) to find out more

NTCIR

- first eval campaign to have a patent-related task in 2002
  - test collection[s]
    - 2 years full text JP
    - 5 years abstracts JP and 5 years abstracts EN
    - topics created manually from news articles
      - all in 5 languages (JP, EN, KO, trad/simplified CN)
      - 6 for training and 25 for testing
    - graded relevance (4 levels)
- 2003/2004 – first invalidity search campaign (similar to Prior Art)
  - results had to order passages of the document in order of importance to the query
  - human evaluations again
    - 7 train topics, 103 test topics (34 manually evaluated, 69 based on the search report)
    - topic – JP patent application rejected by the JPO
    - topics translated into EN and simplified CN

NTCIR

- first eval campaign to have a patent-related task in 2002
  - test collection[s]
    - 2 years full text JP
    - 5 years abstracts JP and 5 years abstracts EN

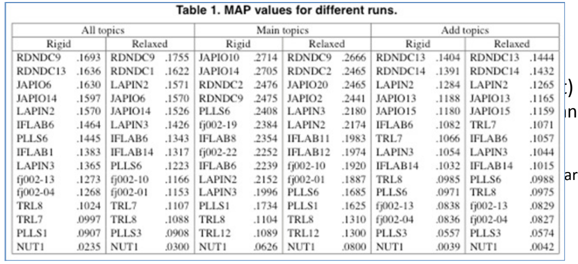


Table 1. MAP values for different runs.

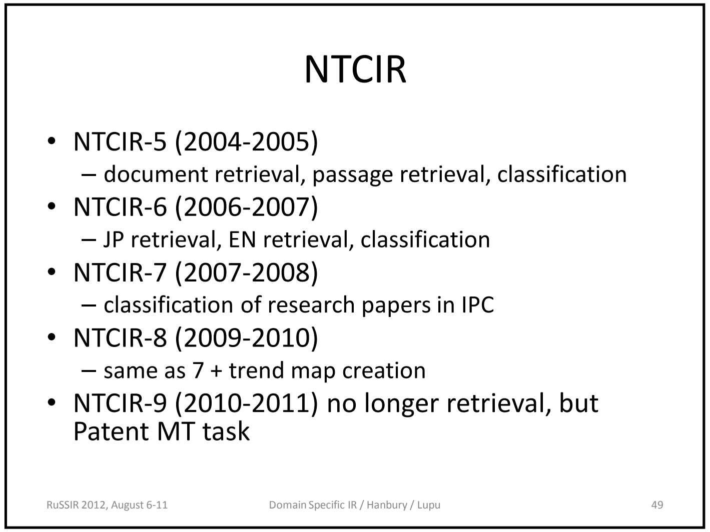
All topics				Main topics				Add topics			
Rigid		Relaxed		Rigid		Relaxed		Rigid		Relaxed	
RDNDC9	.1693	RDNDC9	.1755	JAPIO10	.2714	RDNDC9	.2666	RDNDC13	.1404	RDNDC13	.1444
RDNDC13	.1636	RDNDC1	.1622	JAPIO14	.2705	RDNDC2	.2465	RDNDC14	.1391	RDNDC14	.1432
JAPIO6	.1630	LAPIN2	.1571	RDNDC2	.2476	JAPIO20	.2465	LAPIN2	.1284	LAPIN2	.1265
JAPIO14	.1597	JAPIO6	.1570	RDNDC9	.2475	JAPIO2	.2441	JAPIO13	.1188	JAPIO13	.1165
LAPIN2	.1570	JAPIO14	.1526	PLLS6	.2408	LAPIN3	.2180	JAPIO15	.1180	JAPIO15	.1159
IFLAB6	.1464	LAPIN3	.1426	fj002-19	.2384	LAPIN2	.2174	IFLAB6	.1082	TRL7	.1071
PLLS6	.1445	IFLAB6	.1343	IFLAB8	.2354	IFLAB11	.1983	TRL7	.1066	IFLAB6	.1057
IFLAB1	.1383	IFLAB14	.1317	fj002-22	.2252	IFLAB12	.1974	LAPIN3	.1054	LAPIN3	.1044
LAPIN3	.1365	PLLS6	.1223	IFLAB6	.2239	fj002-10	.1920	IFLAB14	.1032	IFLAB14	.1015
fj002-13	.1273	fj002-10	.1166	LAPIN2	.2152	fj002-01	.1887	TRL8	.0985	PLLS6	.0988
fj002-04	.1268	fj002-01	.1153	LAPIN3	.1996	PLLS6	.1685	PLLS6	.0971	TRL8	.0975
TRL8	.1024	TRL7	.1107	PLLS1	.1734	PLLS1	.1625	fj002-13	.0838	fj002-13	.0829
TRL7	.0997	TRL8	.1088	TRL8	.1104	TRL8	.1310	fj002-04	.0836	fj002-04	.0827
PLLS1	.0907	PLLS3	.0908	TRL12	.1089	TRL12	.1300	PLLS3	.0557	PLLS3	.0574
NUT1	.0235	NUT1	.0300	NUT1	.0626	NUT1	.0800	NUT1	.0039	NUT1	.0042

-

t)

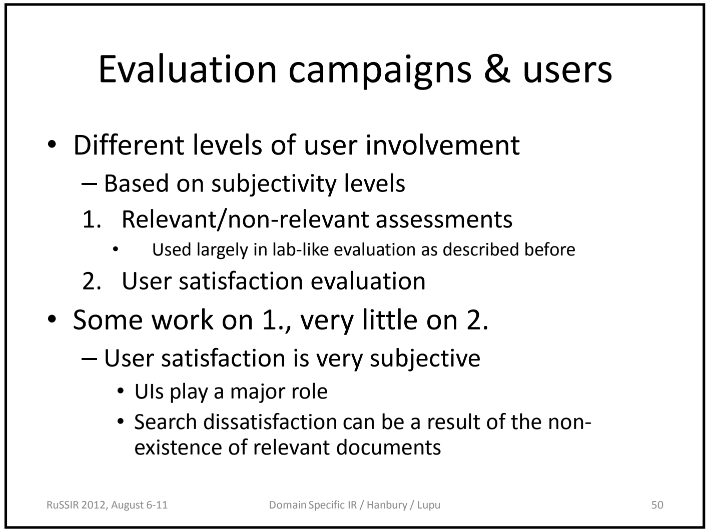
nce

arch



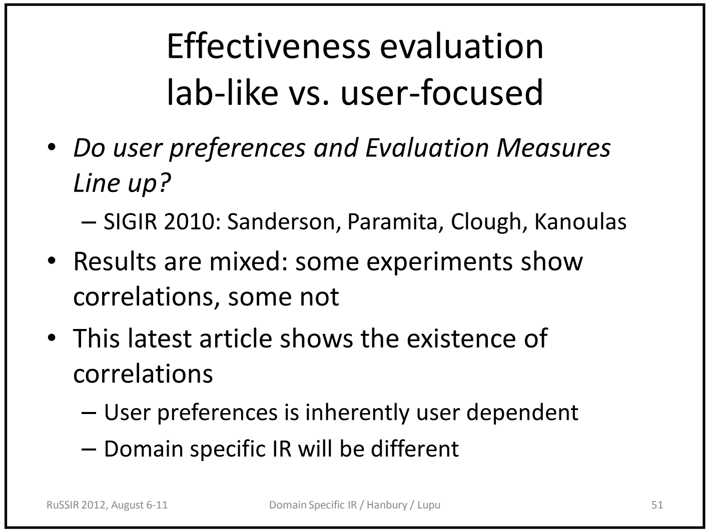
NTCIR

- NTCIR-5 (2004-2005)
  - document retrieval, passage retrieval, classification
- NTCIR-6 (2006-2007)
  - JP retrieval, EN retrieval, classification
- NTCIR-7 (2007-2008)
  - classification of research papers in IPC
- NTCIR-8 (2009-2010)
  - same as 7 + trend map creation
- NTCIR-9 (2010-2011) no longer retrieval, but Patent MT task



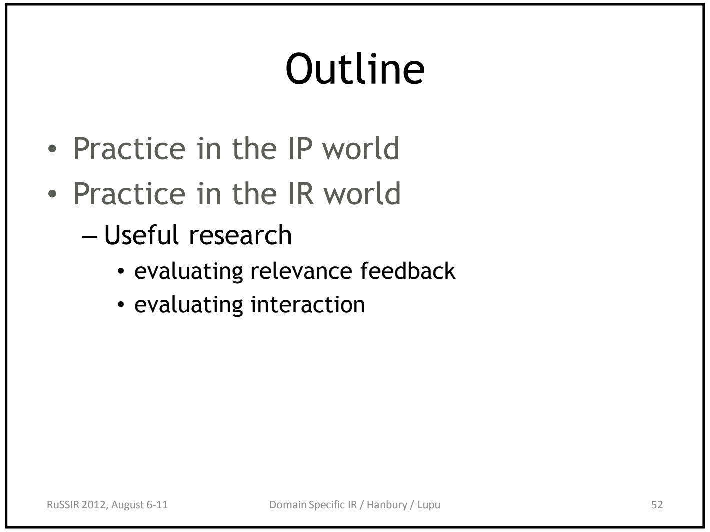
Evaluation campaigns & users

- Different levels of user involvement
  - Based on subjectivity levels
  1. Relevant/non-relevant assessments
     - Used largely in lab-like evaluation as described before
  2. User satisfaction evaluation
- Some work on 1., very little on 2.
  - User satisfaction is very subjective
    - UIs play a major role
    - Search dissatisfaction can be a result of the non-existence of relevant documents



Effectiveness evaluation lab-like vs. user-focused

- *Do user preferences and Evaluation Measures Line up?*
  - SIGIR 2010: Sanderson, Paramita, Clough, Kanoulas
- Results are mixed: some experiments show correlations, some not
- This latest article shows the existence of correlations
  - User preferences is inherently user dependent
  - Domain specific IR will be different



Outline

- Practice in the IP world
- Practice in the IR world
  - Useful research
    - evaluating relevance feedback
    - evaluating interaction

Relevance feedback evaluation

- [Chang:1971] – evaluation of RF algorithms is a problem for precision and recall
  - tendency to just put to the top of the list the documents indicated as relevant
- compensation measures
  - Residual ranking: documents used in RF are removed from the collection
    - + considers only the effect of feedback on the unseen relevant documents
    - - test collection changing -> results not comparable

(cont.) compensation measures

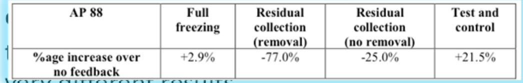
- **freezing** : the top n documents, used to modify the query, are frozen in place
  - + comparable results, scores do not change once all relevant documents have been used in RF (reranking of non-relevant ones only)
  - – scores may decrease as the iterations increase, because non-relevant documents are frozen in place, even though more relevant documents are found
- **test and control groups** : split the collection in two. Query modification is performed by RF on the test group and the new query is then run agains the control group. RP evaluation is only done on the control group, which is free to move in the ranking as needed.
  - + comparable results, freedom of movement
  - – splitting the collection is difficult to do in a sensible manner.

Relevance feedback evaluation

- Problems
  - RF/Interactive IR is is modelling a user who may, over time, change its information need
  - the different compensation measures can give very different results
    - are calculating different aspects of feedback:
      - freezing is measuring cumulative effectiveness,
      - residual collection is measuring the effectiveness of retrieving only the remaining relevant documents,
      - test and control is measuring the relative performance of the modified queries produced at each iteration.

Relevance feedback evaluation

- Problems
  - RF/Interactive IR is is modelling a user who may,



AP 88	Full freezing	Residual collection (removal)	Residual collection (no removal)	Test and control
%age increase over no feedback	+2.9%	-77.0%	-25.0%	+21.5%

very different results

© Ruthven, Lalmas, *A survey of the use of relevance feedback for information access systems. The Knowledge Engineering Review 2003*

- freezing is measuring cumulative effectiveness,
- residual collection is measuring the effectiveness of retrieving only the remaining relevant documents,
- test and control is measuring the relative performance of the modified queries produced at each iteration.

RF evaluation campaign

- TREC Relevance Feedback track
  - from 2008 to 2010
    - 2008 concentrated on the algorithm itself
      - participants were given the same sets of judged docs and used their own algorithms to retrieve new docs
    - 2009 concentrated on finding good sets of docs to base their retrieval on
      - each participant submitted one or two sets of 5 documents for each topic, 3-5 other participants ran with those docs → get a system independent score of how good those docs were
    - 2010 focuses even more, on 1 document only (how good it is for RF)

Session-based IR evaluation

- first organized in 2010

> “A search engine may be able to better serve a user not by ranking the most relevant results to each query in the sequence, but by ranking results that help “point the way” to what the user is really looking for, or by complementing results from previous queries in the sequence with new results, or in other currently-unanticipated ways.”

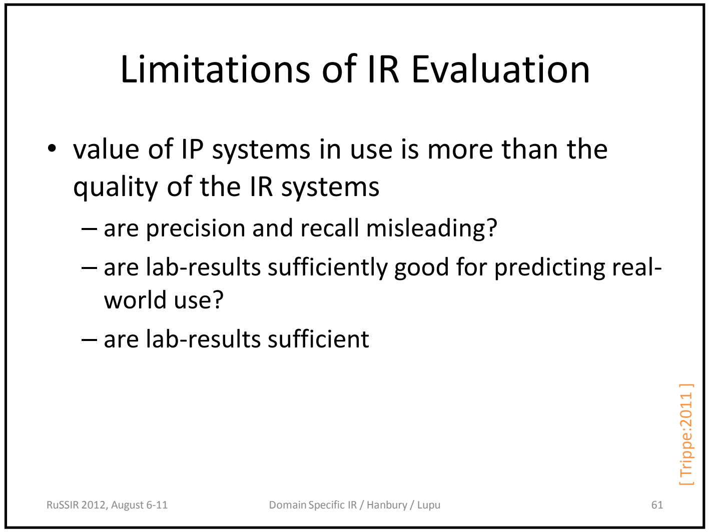
- Objectives
  1. to see if a system can improve results based on knowledge of a previous query
  2. to evaluate system performance over an entire session rather than the usual ad-hoc methodology

Session-based IR evaluation

- 150 query pairs
  - original query : reformulated query
  - three types of reformulations
    - specifications
    - generalizations
    - drifting/parallel reformulations
- for each query, participants submit 3 ranked lists:
  1. over the original query
  2. over the reformulated query only
  3. over the reformulated query taking into account the original one

Other useful research

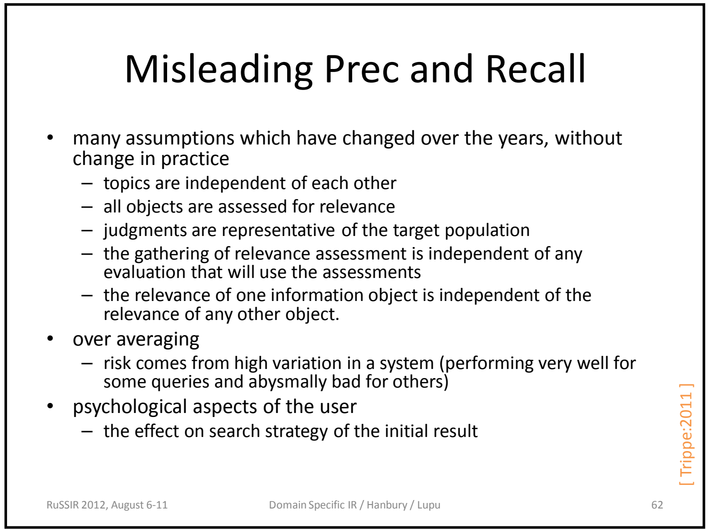
- Retrievability
  - [Azzopardi:2008],[Bashir:2010]
  - because patent search is recall oriented but recall is impossible to compute
  - measure how ‘accessible’/’retrievable’ documents are on random queries
  - objective of an IR systems: have a uniform distribution of retrievability
    - have no documents which are impossible to retrieve
  - [Bashir:2010] shows that pseudo-relevance feedback can significantly skew retrievability



Limitations of IR Evaluation

- value of IP systems in use is more than the quality of the IR systems
  - are precision and recall misleading?
  - are lab-results sufficiently good for predicting real-world use?
  - are lab-results sufficient

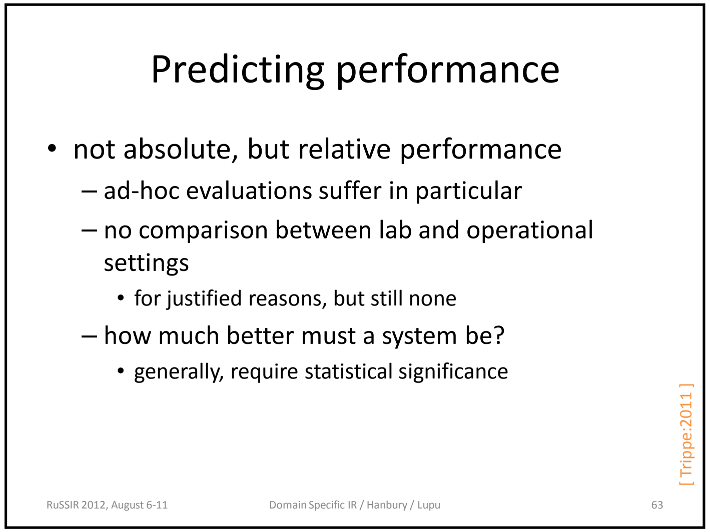
[Trippe:2011]



Misleading Prec and Recall

- many assumptions which have changed over the years, without change in practice
  - topics are independent of each other
  - all objects are assessed for relevance
  - judgments are representative of the target population
  - the gathering of relevance assessment is independent of any evaluation that will use the assessments
  - the relevance of one information object is independent of the relevance of any other object.
- over averaging
  - risk comes from high variation in a system (performing very well for some queries and abysmally bad for others)
- psychological aspects of the user
  - the effect on search strategy of the initial result

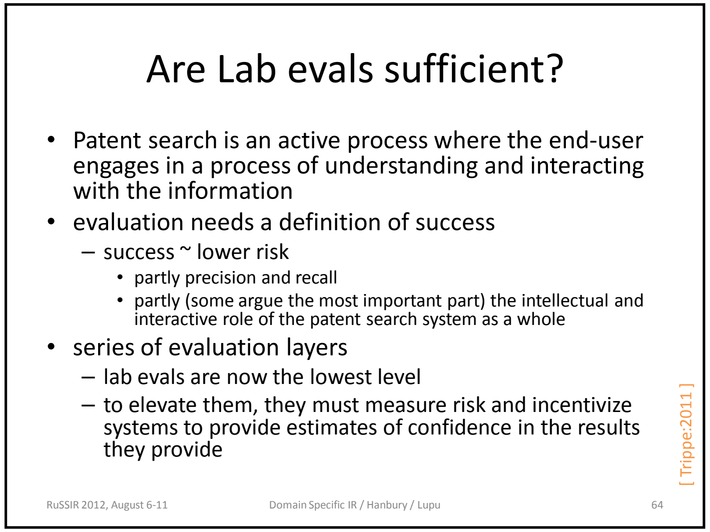
[Trippe:2011]



Predicting performance

- not absolute, but relative performance
  - ad-hoc evaluations suffer in particular
  - no comparison between lab and operational settings
    - for justified reasons, but still none
  - how much better must a system be?
    - generally, require statistical significance

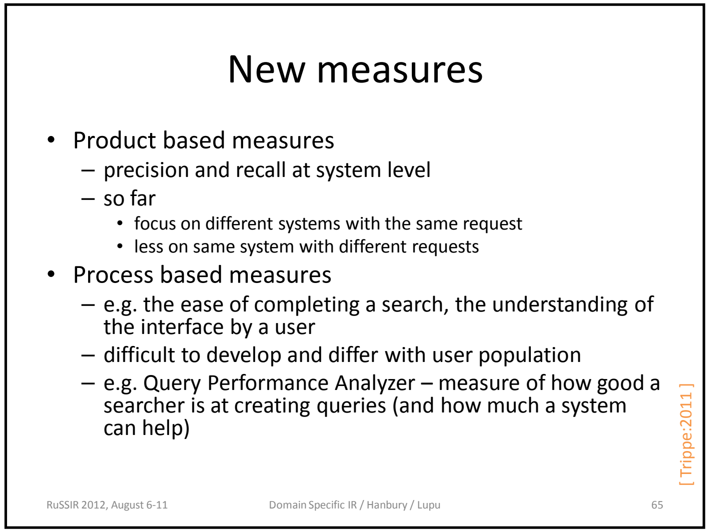
[Trippe:2011]



Are Lab evals sufficient?

- Patent search is an active process where the end-user engages in a process of understanding and interacting with the information
- evaluation needs a definition of success
  - success ~ lower risk
    - partly precision and recall
    - partly (some argue the most important part) the intellectual and interactive role of the patent search system as a whole
- series of evaluation layers
  - lab evals are now the lowest level
  - to elevate them, they must measure risk and incentivize systems to provide estimates of confidence in the results they provide

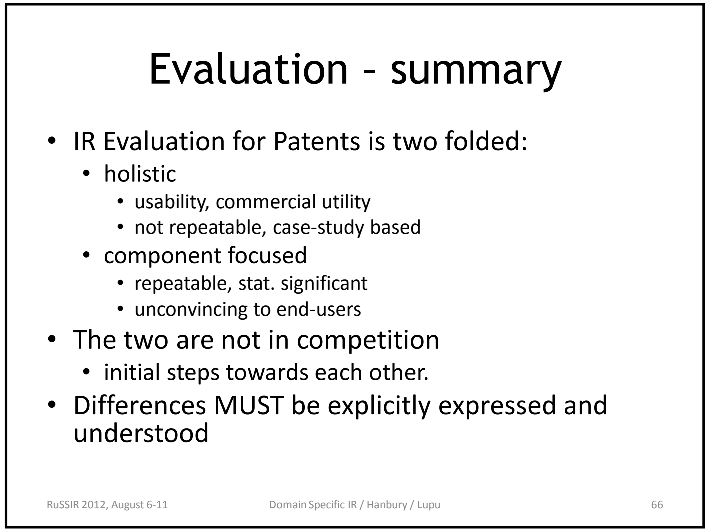
[Trippe:2011]



New measures

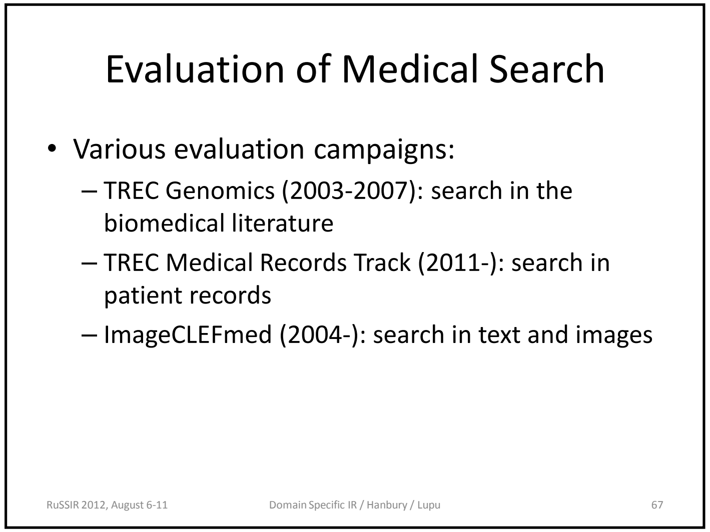
- Product based measures
  - precision and recall at system level
  - so far
    - focus on different systems with the same request
    - less on same system with different requests
- Process based measures
  - e.g. the ease of completing a search, the understanding of the interface by a user
  - difficult to develop and differ with user population
  - e.g. Query Performance Analyzer – measure of how good a searcher is at creating queries (and how much a system can help)

[Trippe:2011]



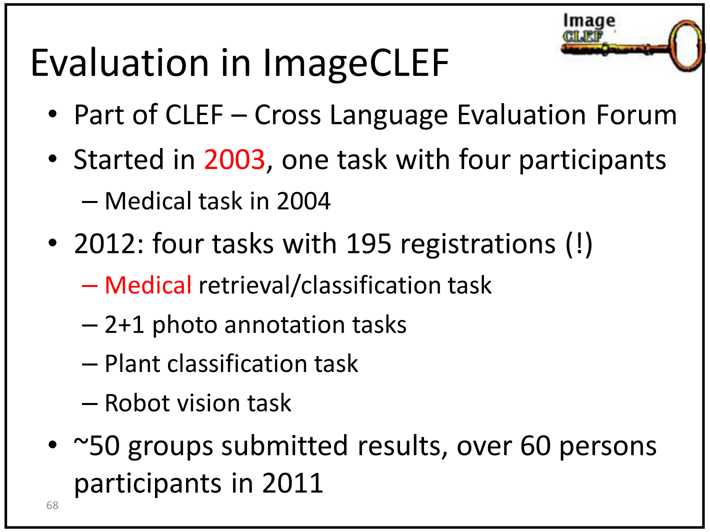
Evaluation - summary

- IR Evaluation for Patents is two folded:
  - holistic
    - usability, commercial utility
    - not repeatable, case-study based
  - component focused
    - repeatable, stat. significant
    - unconvincing to end-users
- The two are not in competition
  - initial steps towards each other.
- Differences MUST be explicitly expressed and understood



Evaluation of Medical Search

- Various evaluation campaigns:
  - TREC Genomics (2003-2007): search in the biomedical literature
  - TREC Medical Records Track (2011-): search in patient records
  - ImageCLEFmed (2004-): search in text and images



Evaluation in ImageCLEF

- Part of CLEF – Cross Language Evaluation Forum
- Started in 2003, one task with four participants
  - Medical task in 2004
- 2012: four tasks with 195 registrations (!)
  - Medical retrieval/classification task
  - 2+1 photo annotation tasks
  - Plant classification task
  - Robot vision task
- ~50 groups submitted results, over 60 persons participants in 2011

Image databases

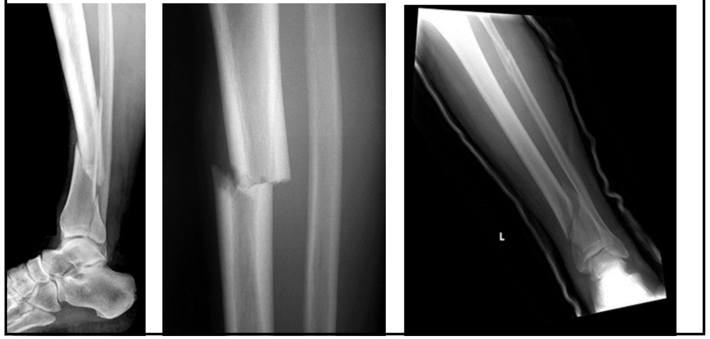
- Databases should change every 2-4 years
  - Or impact declines (see ImageCLEF impact analysis)
- Increasingly large
  - 8000 images in 2004, 300 000 in 2012 for the medical task
- Copyright problems are often present
  - Now images taken from the open access literature
    - Frequent in the medical domain (PubMedCentral 1.5 Mio)
  - Redistribution not necessarily possible
- Databases need to have challenges
  - Irrelevant images, stock photography, …

Tasks for image retrieval

- What are realistic tasks to compare systems upon?
- Possibilities to define tasks
  - Survey among end users
    - 5 surveys performed so far for the medical task
      - Do people know what systems can do?
    - Khresmoi project aims also at lay persons and GPs
  - Analysis of related log files
    - Are there any visual search systems?
    - MedLine, HON Media search, Goldminer
- Multilingual and multimodal topics

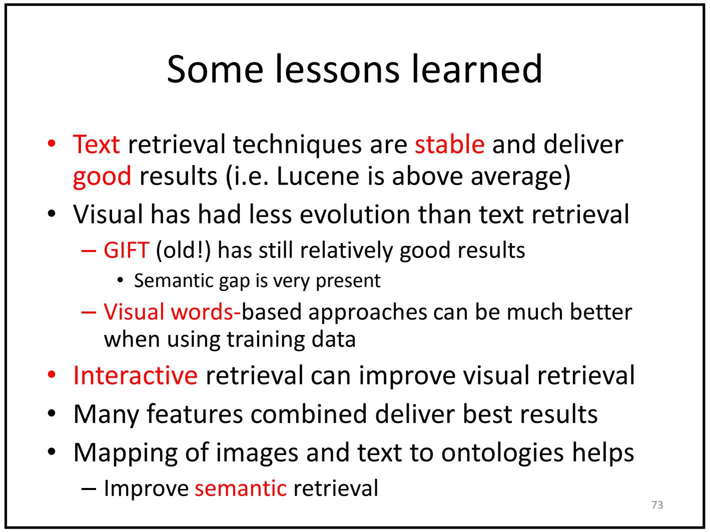
Example task

- Show me x-ray images of a tibia with a fracture.
- Zeige mir Röntgenbilder einer gebrochenen Tibia.
- Montre-moi des radiographies du tibia avec fracture.



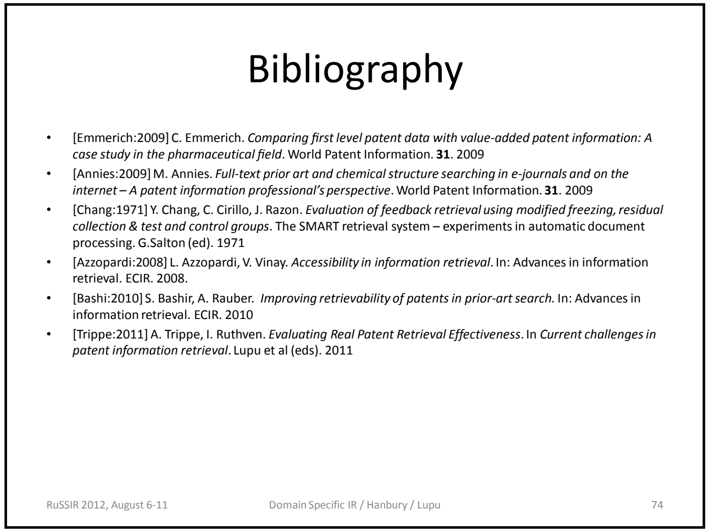
Ground truthing and performance measures

- For medical task MDs judge all images (plus double judgments for consistency)
- Measures can be heavily debated
  - MAP, early precision, Bpref, …
  - What would be most user-oriented?
  - Accuracy for classification? Specificity and sensitivity?
- Most often pooling is used
  - Not all images are judged for relevance
  - Ternary judgment scheme (relevant, non-relevant, partly relevant)



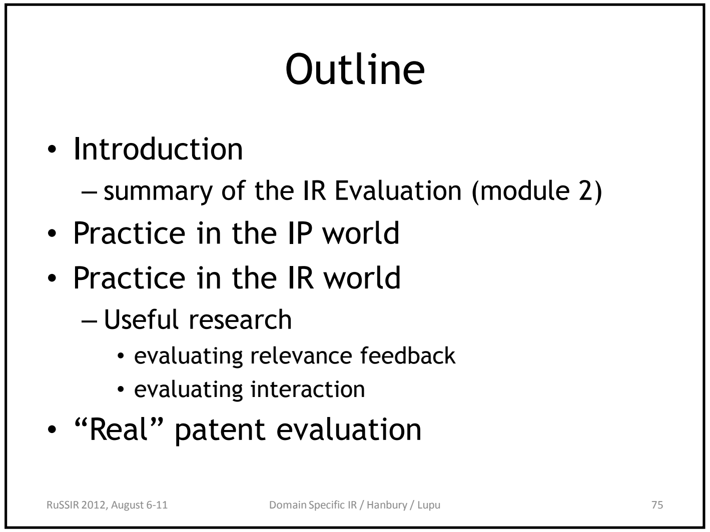
Some lessons learned

- Text retrieval techniques are stable and deliver good results (i.e. Lucene is above average)
- Visual has had less evolution than text retrieval
  - GIFT (old!) has still relatively good results
    - Semantic gap is very present
  - Visual words-based approaches can be much better when using training data
- Interactive retrieval can improve visual retrieval
- Many features combined deliver best results
- Mapping of images and text to ontologies helps
  - Improve semantic retrieval



Bibliography

- [Emmerich:2009] C. Emmerich. *Comparing first level patent data with value-added patent information: A case study in the pharmaceutical field*. World Patent Information. **31**. 2009
- [Annies:2009] M. Annies. *Full-text prior art and chemical structure searching in e-journals and on the internet – A patent information professional's perspective*. World Patent Information. **31**. 2009
- [Chang:1971] Y. Chang, C. Cirillo, J. Razon. *Evaluation of feedback retrieval using modified freezing, residual collection & test and control groups*. The SMART retrieval system – experiments in automatic document processing. G.Salton (ed). 1971
- [Azzopardi:2008] L. Azzopardi, V. Vinay. *Accessibility in information retrieval*. In: Advances in information retrieval. ECIR. 2008.
- [Bashi:2010] S. Bashir, A. Rauber. *Improving retrievability of patents in prior-art search*. In: Advances in information retrieval. ECIR. 2010
- [Trippe:2011] A. Trippe, I. Ruthven. *Evaluating Real Patent Retrieval Effectiveness*. In *Current challenges in patent information retrieval*. Lupu et al (eds). 2011



Outline

- Introduction
  - summary of the IR Evaluation (module 2)
- Practice in the IP world
- Practice in the IR world
  - Useful research
    - evaluating relevance feedback
    - evaluating interaction
- "Real" patent evaluation