Multimodal Information Retrieval Evaluation

Allan Hanbury

Outline

- Introduction
- Retrieval effectiveness evaluation
- Evaluation campaigns
- User-based evaluation
- Conclusion

paris - Google Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.google.co.uk/search?hl=en&q=paris&btnG=Search&meta=

Most Visited Getting Started Latest Headlines SAP NetWeaver Portal

Web Images Maps News Shopping Google Mail more ▾ Sign in

Google paris Search Advanced Search Preferences

Search: the web pages from the UK

Web Results **1** - **10** of about **527,000,000** for **paris** [definition]. (**0.18** seconds)

Sponsored Links

Paris
www.expedia.co.uk Tailor make a perfect holiday trip to **Paris** only at Expedia!

Paris must sees
www.disneylandparis.com Don't miss Disneyland, 35 mins from **Paris**. Book now online and save 15%

Paris
France
maps.google.co.uk

Paris - Wikipedia, the free encyclopedia
It is situated on the river Seine, in northern France, at the heart of the Île-de-France region (also known as the "**Paris** Region"; French: Région ...
en.wikipedia.org/wiki/**Paris** - 409k - Cached - Similar pages

The **Paris** Pages / Les Pages de **Paris** / **Paris**.Org / The Web's First ...
Org ™ / Online Since 14 July 1994, It's The Internet's First **Paris** Web Site. ... Get on and off along the tour route and see **Paris** at your own pace. ...
www.**paris**.org/ - 23k - Cached - Similar pages

All about **Paris**: Hotels, City Guide, Tours, Airport Shuttle ...
All about **Paris**: Hotels, City Guide, Tours, Airport Shuttle and Restaurants.
paris.com/ - 11k - Cached - Similar pages

Sponsored Links

Flights to **Paris** Sale
Up to 1/3 off all routes and hotels
Offer ends 17th Feb. Book Now!
www.bmibaby.com

Paris Eurostar Breaks
Short Or Luxury Trips To **Paris** With Eurostar, See Our Best Deals Online
www.railbookers.com

Over 350 Hotels in **Paris**
Save up to 75% on your booking.
Low rates and great availability!
www.booking.com

Paris Holidays
Fantastic holidays to **Paris**
Low deposit of only £30 per person!
www.thomson.co.uk

Hotels in **Paris**
400 Hotels - Up to 60% Discount!
Charm, Design, Luxury or Family
www.FastBooking.com

Paris
Read the expert guide on where to go and what to do in **Paris**.
www.cntraveller.co.uk

Done

paris - Google Image Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://images.google.co.uk/images?um=1&hl=en&safe=off&q=paris&btnG=Search+In

Most Visited Getting Started Latest Headlines SAP NetWeaver Portal

Web **Images** Maps News Shopping Google Mail more ▾ Sign in

Google paris Search Images Search the Web Advanced Image Search Preferences

SafeSearch is off

Images Showing: All image sizes Any content Results **1** - **20** of about **49,400,000** for **paris** [definition]. (**0.04** seconds)

Related searches: paris **france** paris **map** paris **city** **eiffel tower**

Paris South Suburbs
560 x 418 - 45k - jpg
www.a-t-s.net

Anyone who visits **Paris** has to stand ...
375 x 500 - 65k - jpg
cruises.about.com

Paris Hilton punts burger
300 x 425 - 41k - jpg
www.theregister.co.uk

Day 1-**Paris** · Day 2-**Paris**
450 x 619 - 43k - jpg
www.wired2theworld.com

Paris Hilton feels her boobs slowly ...
413 x 459 - 57k - jpg
www.bestweekever.tv

Paris-1.jpg
600 x 386 - 67k - jpg
www.airportdirecttravel.co.uk

From: **Paris**
550 x 412 - 36k - jpg
www.tripadvisor.com
[More from media-cdn.tripadvisor.com]

Study Abroad in **Paris**.
300 x 450 - 40k - jpg
www.iiepassport.org

Done

paris france - Google Image Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://images.google.co.uk/images?um=1&hl=en&safe=off&q=paris+france&btnG=Search+Images

Most Visited Getting Started Latest Headlines SAP NetWeaver Portal

Web **Images** Maps News Shopping Google Mail more ▼ Sign in

Google paris france Search Images Search the Web Advanced Image Search Preferences

SafeSearch is off

Images Showing: All image sizes Any content Results **1** - **20** of about **19,300,000** for **paris** **france**. (**0.03** seconds)

Related searches: **eiffel tower**

australia **paris** england
617 x 896 - 262k - jpg
www.uiu.edu

France ...
415 x 332 - 28k - jpg
www.destination360.com
[More from www.destination360.com]

Paris, **France**, View of the Eiffel ...
339 x 450 - 45k
www.allposters.com

Arc de Triomphe, **Paris**, **France**.
360 x 450 - 53k - jpg
www.britannica.com

Louvre, **Paris**, **France**.
1600 x 1200 - 482k
www.photos4travel.com

Picture of The Louvre, **Paris**, **France ...**
600 x 400 - 114k
www.freefoto.com
[More from www.freefoto.com]

Paris France ...
415 x 332 - 48k - jpg
www.destination360.com

Paris France Fireworks Eiffel tower ...
667 x 1000 - 170k
www.fromparis.com

Paris, **France** real estate market ...
505 x 311 - 42k - jpg
matrix.millersamuel.com

... Eiffel Tower **Paris France**
1024 x 768 - 130k - jpg
www.fantom-xp.com

Done

paris -hilton - Google Image Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://images.google.co.uk/images?gbv=2&hl=en&safe=off&q=paris+-hilton&btnG=Search+Images

Most Visited Getting Started Latest Headlines SAP NetWeaver Portal

Web **Images** Maps News Shopping Google Mail more ▾ Sign in

Google paris -hilton Search Images Search the Web Advanced Image Search Preferences

SafeSearch is off

Images Showing: All image sizes Any content Results **1** - **20** of about **38,000,000** for **paris -hilton**. (**0.03** seconds)

My Last Summer in **Paris**: A Little ...
300 x 313 - 15k - jpg
www.gonomad.com

Paris North Suburbs
560 x 418 - 45k - jpg
www.a-t-s.net

Les Escaliers de Montmartre, **Paris** ...
321 x 450 - 56k
www.allposters.com

Eiffel Tower at Night, **Paris**
640 x 480 - 37k - gif
www.afn.org

Day 1-**Paris** · Day 2-**Paris**
450 x 619 - 43k - jpg
www.wired2theworld.com

Paris-1.jpg
600 x 386 - 67k - jpg
www.airportdirecttravel.co.uk

Moulin Rouge, **Paris**
1128 x 758 - 91k - jpg
www.francetourism.com

Paris le
700 x 556 - 58k - jpg
www.e-architect.co.uk

day two in **paris**
2048 x 1536 - 511k - jpg
www.gapingvoid.com

Paris Shopping
415 x 332 - 57k - jpg
www.destination360.com

Done

parijs - Google Image Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://images.google.co.uk/images?um=1&hl=en&safe=off&q=parijs&btnG=Search+Images

Most Visited Getting Started Latest Headlines SAP NetWeaver Portal

Web **Images** Maps News Shopping Google Mail more ▾ Sign in

Google parijs Search Images Search the Web Advanced Image Search Preferences

SafeSearch is off

Images Showing: All image sizes Any content Results **1** - **20** of about **378,000** for **parijs**. (**0.14** seconds)

Parijs
615 x 615 - 322k - jpg
www.lovah.nl

Parijs in beeld december 7, 2006
1600 x 1200 - 446k - jpg
dawnofnone.wordpress.com

Parijs
479 x 502 - 78k - jpg
www.03006.06jn.thinkquest.nl
[More from www.03006.06jn.thinkquest.nl]

Weekendje **Parijs** 2006
1536 x 1024 - 133k - jpg
www.avondschool.be

Parijs
320 x 380 - 25k - jpg
www.zuiden.com
[More from www.zuiden.com]

In 1914 moest in **Parijs** het ...
490 x 368 - 54k - jpg
www.vliegticket4u.nl
[More from www.vliegticket4u.nl]

Parijs
320 x 380 - 23k - jpg
www.zuiden.com

Parijs staat bekend als een van de ...
500 x 500 - 70k - jpg
www.take-a-trip.eu

... van de GWP in **Parijs** en van het ...
399 x 495 - 35k - jpg
www.kawaregem.be

parijs
490 x 490 - 119k - jpg
www.vliegticket4u.nl

Done

Introduction

- Why evaluation?
  - Because without evaluation, there is no research
- Why is this a research field in itself?
  - Because there are many kinds of IR
    - With different evaluation criteria
  - Because it's hard
    - Why?
      - Because it involves human subjectivity (document relevance)
      - Because of the amount of data involved (who can sit down and evaluate 1,750,000 documents returned by Google for 'university vienna'?)

Kinds of 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 $\rightarrow$ more research into it
- And…

Kinds of evaluation (cont.)

- 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)

Outline

- Introduction
- Retrieval effectiveness evaluation
- Evaluation campaigns
- User-based evaluation
- Conclusion

Measures: Precision and Recall

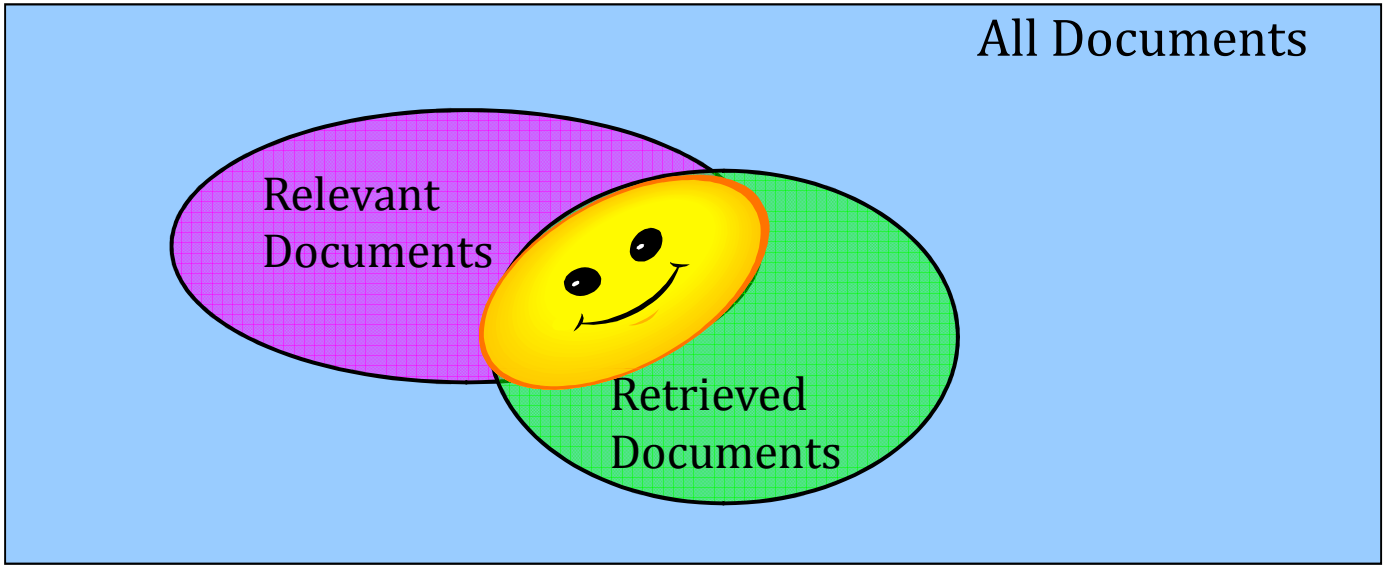
- The search engine returns a list of results
- How do you know how good these results are?
- There are two key concepts in measuring this: precision and recall

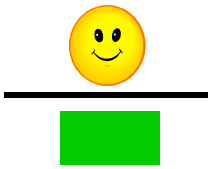
Precision and Recall

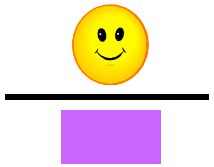
- A query returns *n* ranked documents from a database of many.
- Each one is judged as relevant or not:

Rank	Relevant
1	YES
2	YES
3	NO
4	YES
5	NO
…	
n	NO

Precision and Recall concepts



- Precision = 

Recall = 

Retrieval effectiveness

- Precision
  - How happy are we with what we've got?

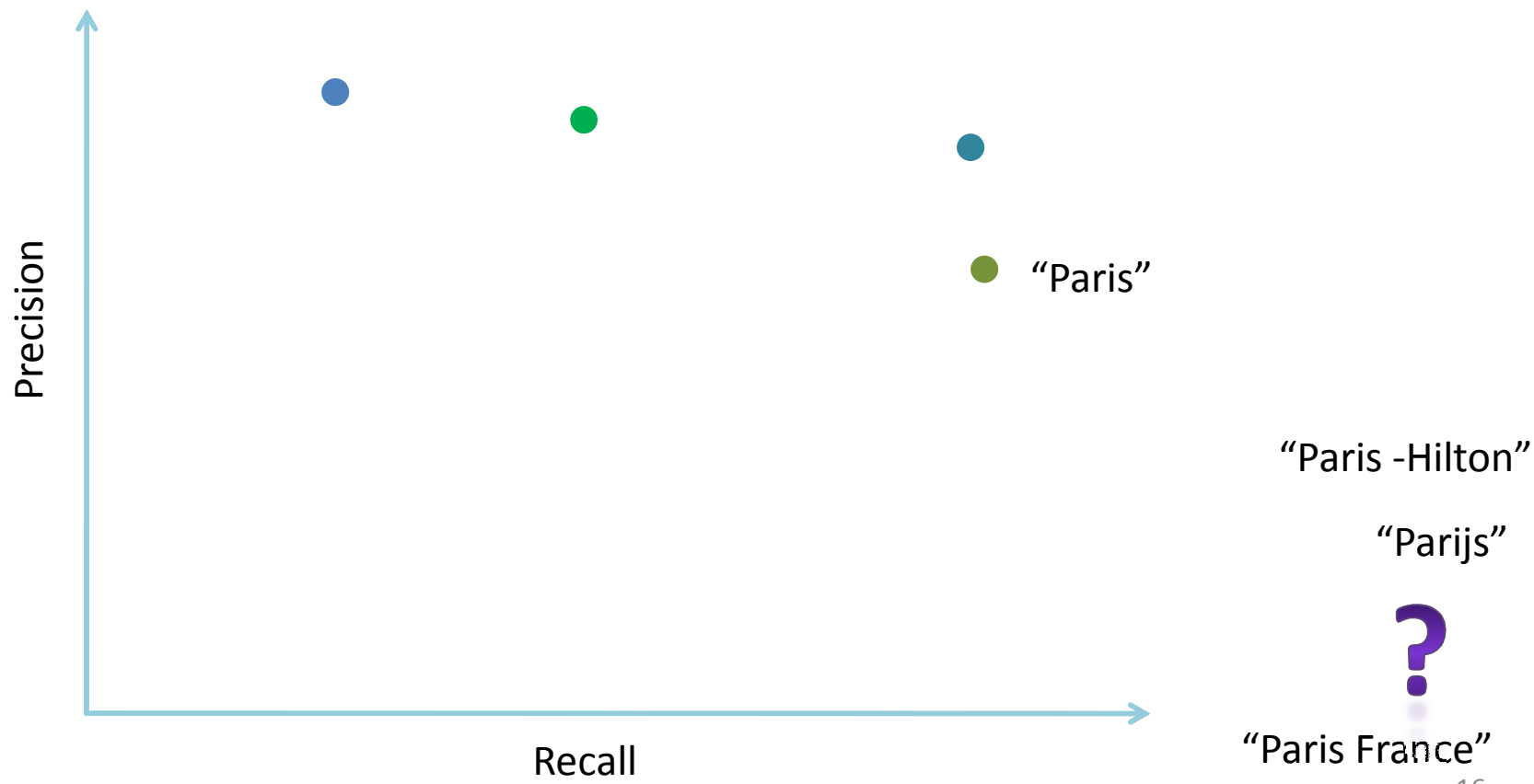
$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}}$$

- Recall
  - How much more could we have had?

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents}}$$

Intuition for Precision and Recall

- Aim is to find all images of the city Paris indexed by Google image search.
- How would you expect precision / recall to behave (roughly)?



Important

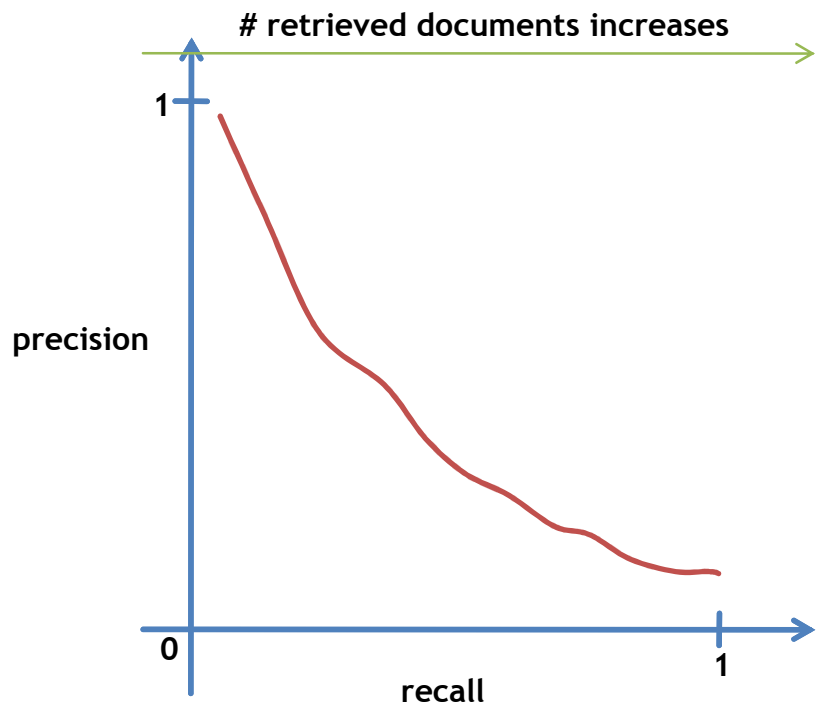
- Quoting Precision or Recall on their own do not make sense.
- How can you obtain a recall of 1.0?
  - Return all the documents in the database.
- How can you make the precision as high as possible?
  - Return only a few documents.

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
- Let’s assume for the moment that’s all we need and that we have it

Retrieval effectiveness

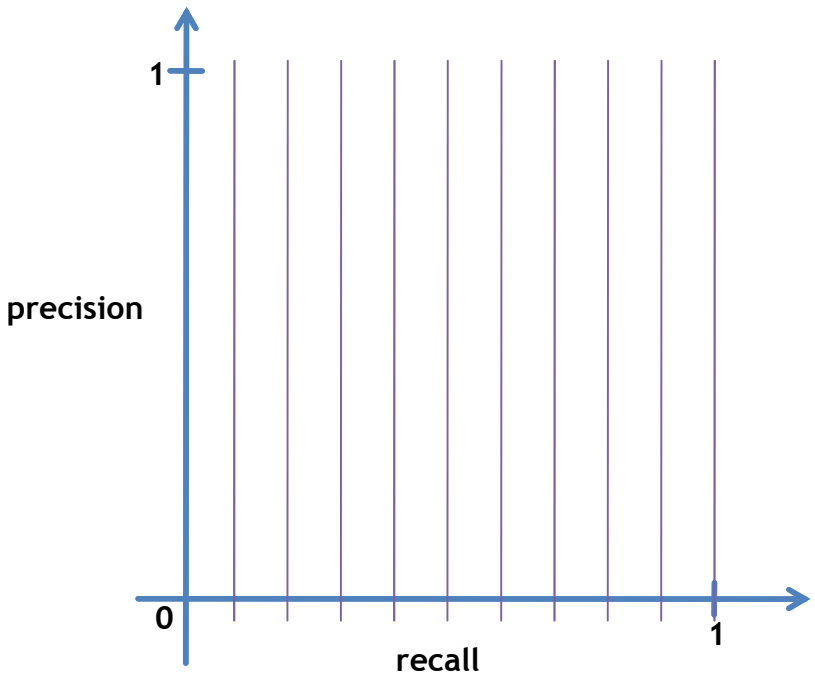
- Precision and Recall generally plotted as a "Precision-Recall curve"



- **They do not play well together**

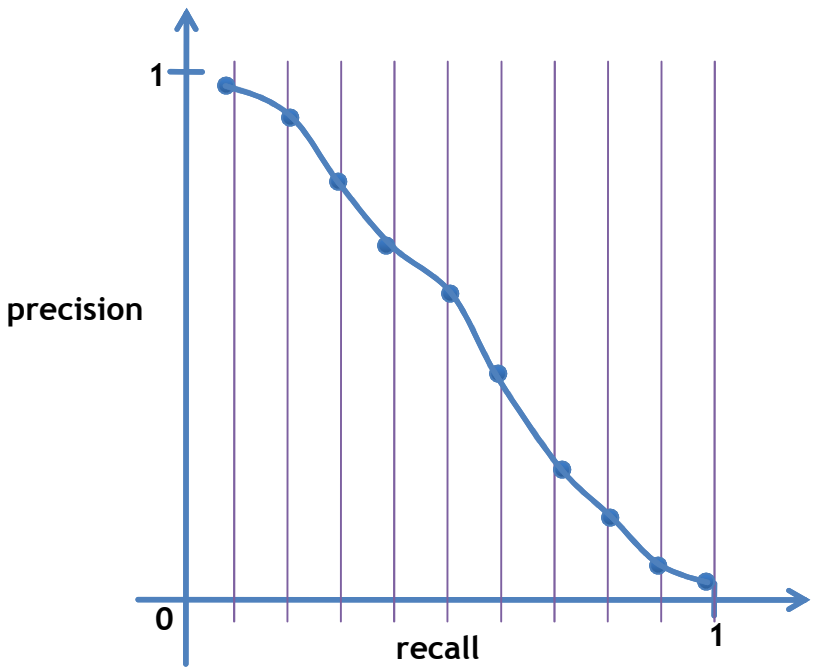
Precision-Recall curves

- How to build a Precision-Recall Curve?
  - For one query at a time
  - Make checkpoints on the recall-axis



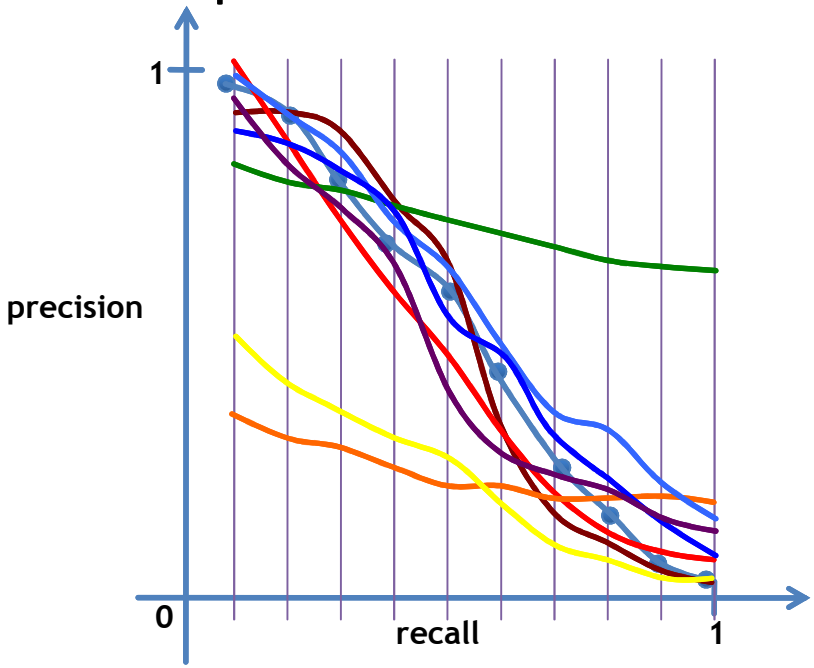
Precision-Recall curves

- How to build a Precision-Recall Curve?
  - For one query at a time
  - Make checkpoints on the recall-axis

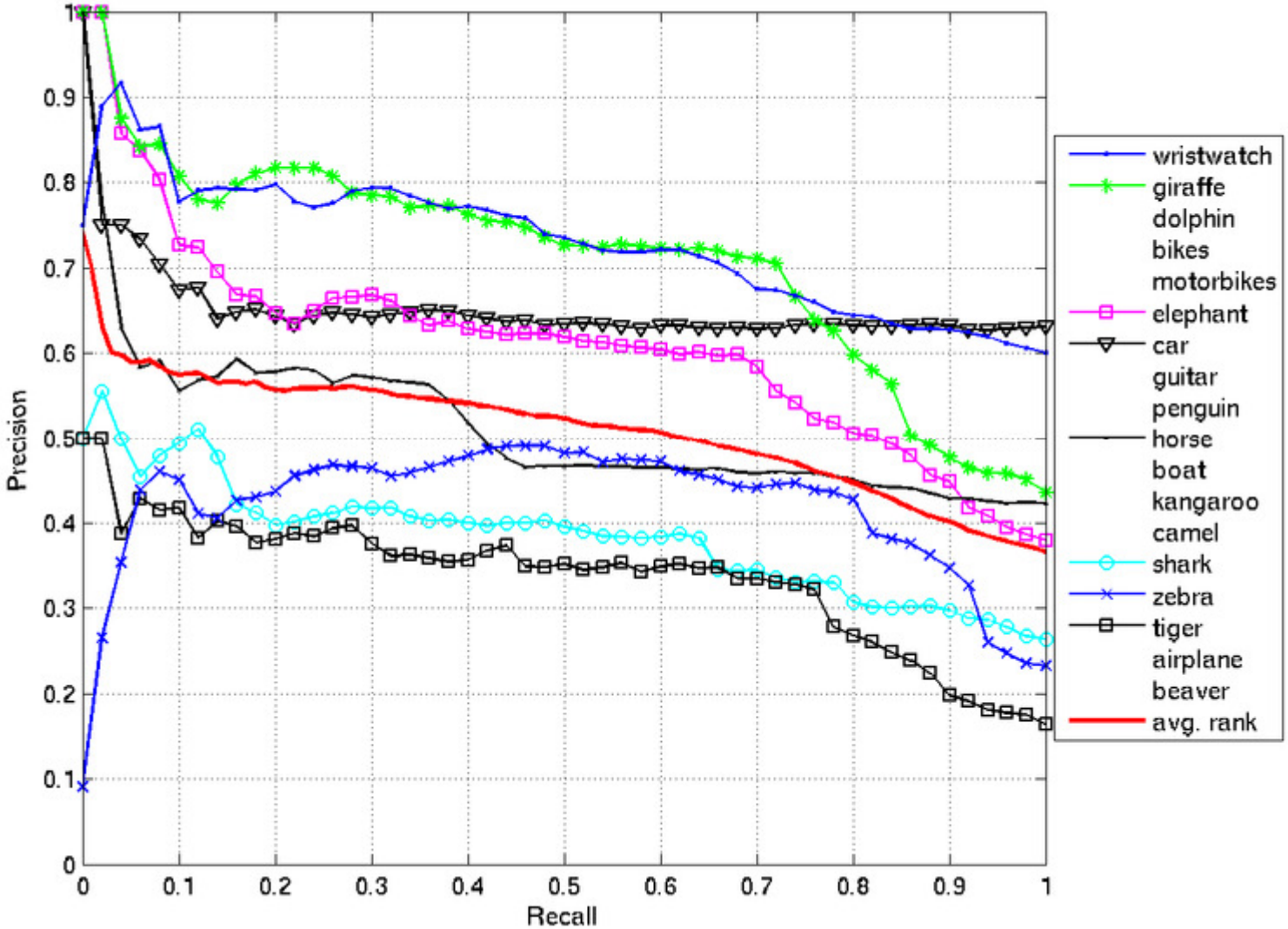


Precision-Recall curves

- How to build a Precision-Recall Curve?
  - For one query at a time
  - Make checkpoints on the recall-axis
  - Repeat for all queries



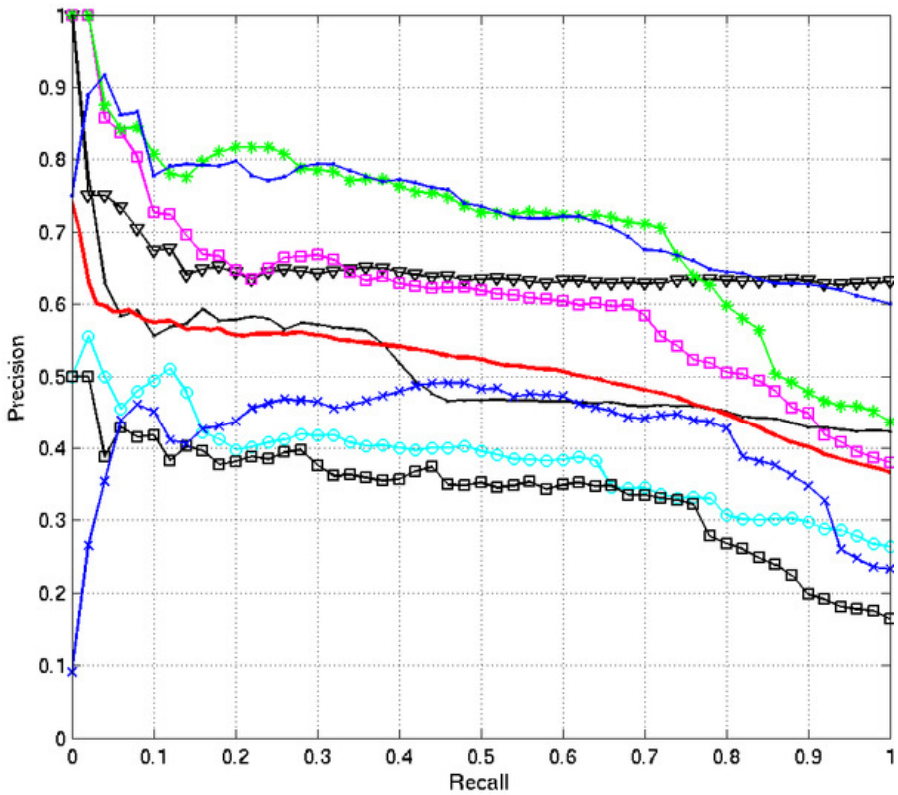
- Examples of Precision-Recall curves



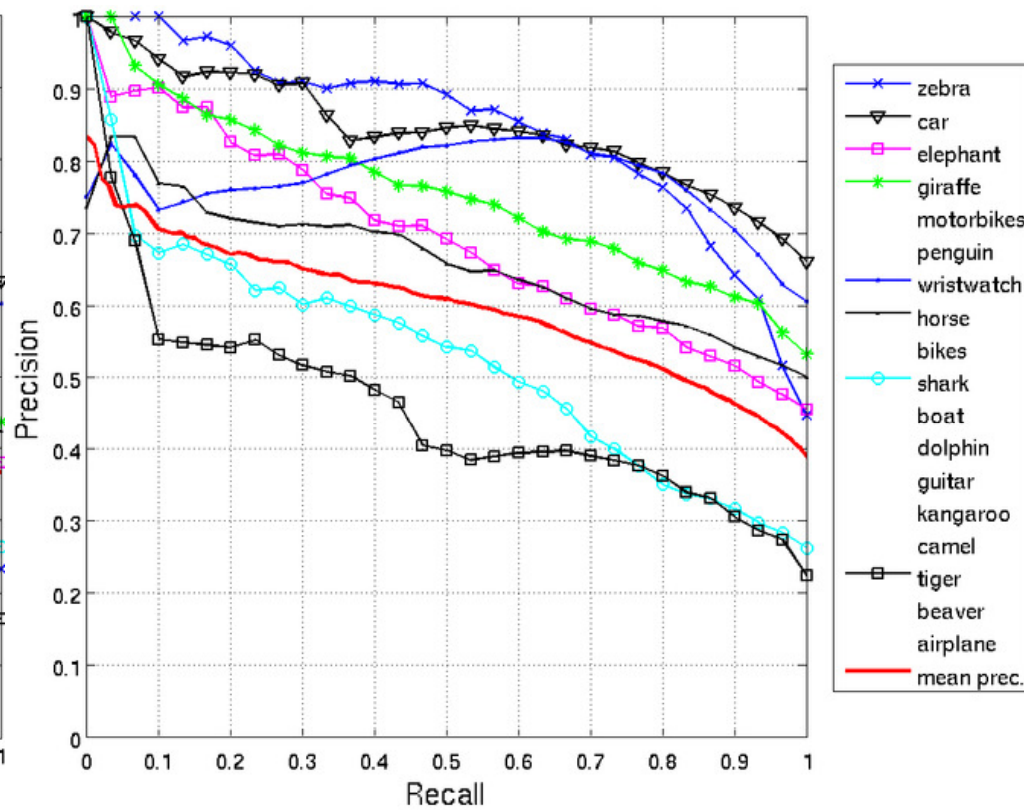
Retrieval of images of a specific object from a database of images.

From: http://www.robots.ox.ac.uk/~vgg/research/mkdb/

- Which results are better?



Text only

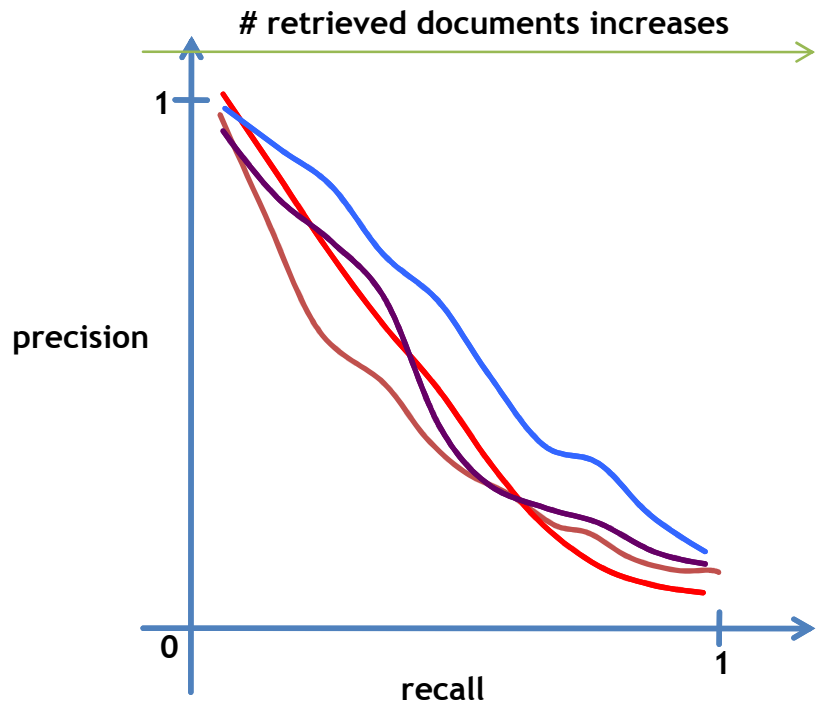


Text and Image Features

From: http://www.robots.ox.ac.uk/~vgg/research/mkdb/

Precision-Recall curves

- The average is the system's P-R curve



- We can compare systems by comparing the curves

Retrieval effectiveness

- What if we don't like this twin-measure approach?
- A solution: F-measure
  - Weighted harmonic mean

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

  - General form for non-negative real β is

$$F_\beta = \frac{\left(1+\beta^2\right) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

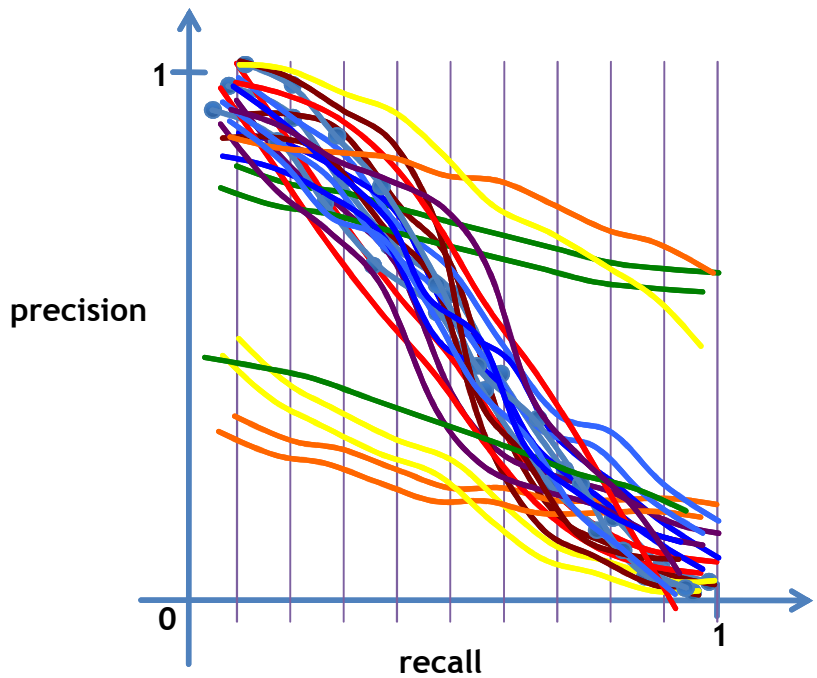
    - F_2 weights recall twice as much as precision
    - $F_{0.5}$ weights precision twice as much as recall

Retrieval effectiveness

- Not quite done yet…
  - 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?

Single-value measures

- What if we want to compare systems at query level?



- Could we have just one measure, to avoid the curves?
  - Note that the F-measure still doesn't solve this (it depends on the cutoff value)

Single-value measures

- Average precision
  - For each query:
    - Every time a relevant document is retrieved, calculate precision
    - Average with previously calculated values
    - Repeat until all relevant documents retrieved
  - For each system:
    - 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.

- $P(n)$ – Precision at n documents
  - Precision when n documents have been retrieved
- Average Precision (AP) emphasizes returning more relevant documents earlier:

rel(r) is 1 if the document at rank r is relevant, 0 otherwise

$$AP = \frac{\sum_{r=1}^{N} [P(r) \times \mathrm{rel}(r)]}{\text{number of relevant documents}}$$

- Mean Average Precision (MAP) is the mean of the AP's for a group of queries

- Example: 4 relevant documents, *N*=5 documents were retrieved:

Rank	Relevant
1	YES
2	YES
3	NO
4	YES
5	NO

Precision =

Recall =

$$AP = \frac{\sum_{r=1}^{N} \left[P(r) \times \mathrm{rel}(r) \right]}{\text{number of relevant documents}}$$

$$AP = \frac{1\times1 + 1\times1 + \frac{2}{3}\times0 + \frac{3}{4}\times1 + \frac{3}{5}\times0}{4} =$$

Retrieval effectiveness

- Not quite done yet…
  - 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?

Cumulative Gain

- For each document *d*, and query *q*, define

  rel(*d*,*q*) >= 0

  - The higher the value, the more relevant the document is to the query
  - Example: (5, 2, 4, 5, 5, 1, 0, 2, 4, …)

$$\mathrm{CG}_p = \sum_{i=1}^{p} rel_i$$

- Pitfalls:
  - Graded relevance introduces even more ambiguity in practice
  - With great flexibility comes great responsibility to justify parameter values

Retrieval effectiveness

- Not quite done yet…
  - 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?

Discounted Cumulative Gain

- A system that returns highly relevant documents at the top of the list should be scored higher than one that returns the same documents lower in the ranked list

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

- Other formulations also possible
- Neither CG, nor DCG can be used for comparison! (depend on # rel documents per query)

Normalised Discounted Cumulative Gain

- Compute DCG for the optimal return set

  E.g.: for a returned set:

  (5,3,5,4,2,0,1,1,5,4,2,2,1,3,3,3,1,0,1,1,0,0..)

  The following:

  (5,5,5,4,4,3,3,3,3,2,2,2,1,1,1,1,1,1,0,0,0,0..)

  has the Ideal Discounted Cumulative Gain: IDCG

- Normalise:

$$\mathrm{nDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p}$$

Other 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…

Other measures

- How about correlations between measures?

	P(30)	R-Prec	MAP	.5 prec	R(1,1000)	Rel Ret	MRR
P(10)	0.88	0.81	0.79	0.78	0.78	0.77	0.77
P(30)		0.87	0.84	0.82	0.80	0.79	0.72
R-Prec			0.93	0.87	0.83	0.83	0.67
MAP				0.88	0.85	0.85	0.64
.5 prec					0.77	0.78	0.63
R(1,1000)						0.92	0.67
Rel ret							0.66

- Kendal Tau values
  - From Voorhees and Harman,2004
- Overall they correlate

Retrieval effectiveness

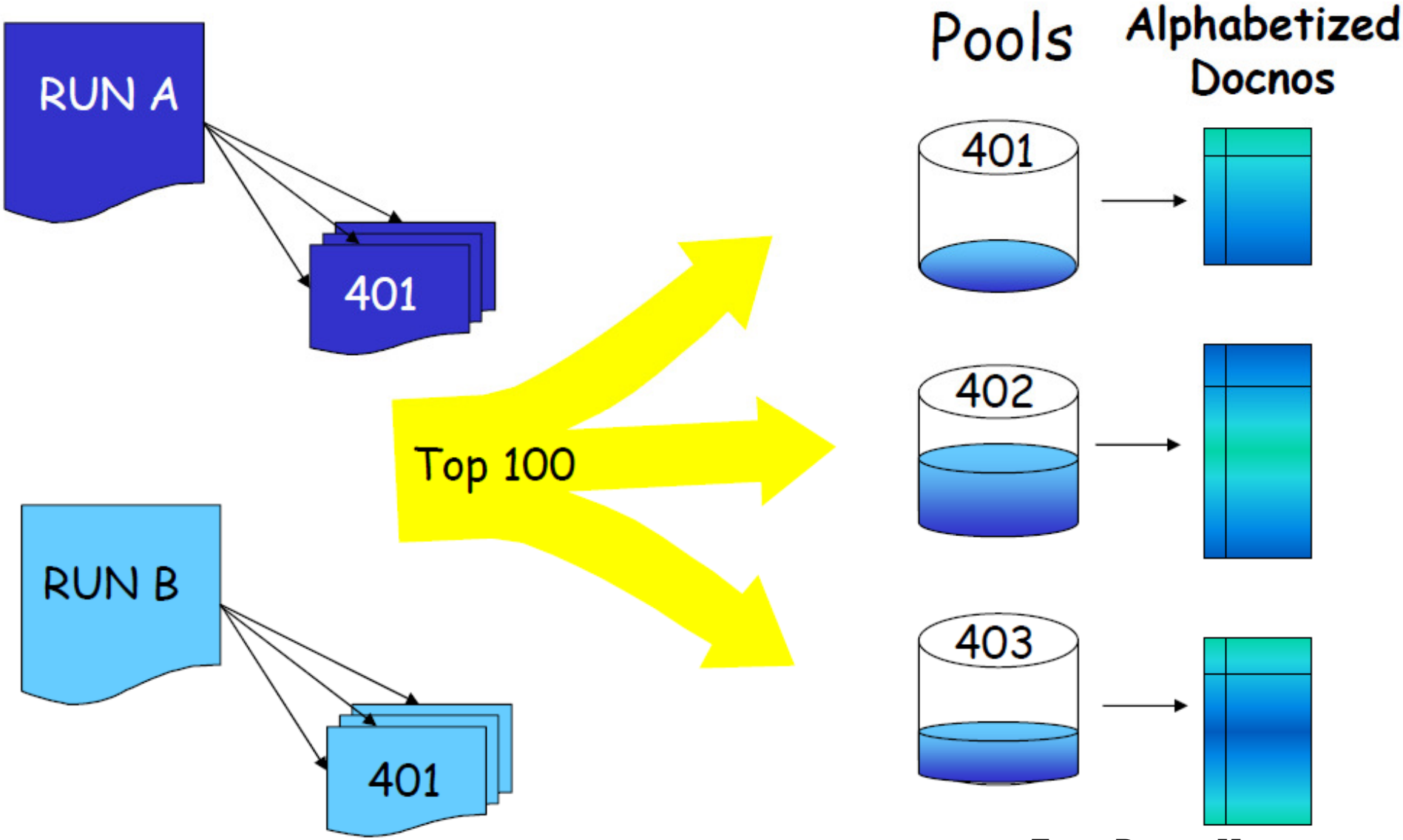
- Not quite done yet…
  - When to stop retrieving?
    - Both P and R imply a cut-off value
  - How about graded relevance
    - Some documents may be more relevant to the question than others
  - How about ranking?
    - A document retrieved at position 1,234,567 can still be considered useful?
  -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

Relevance assessments

- Ideally
  - Sit down and look at all documents
- Practically
  - The ClueWeb09 collection has
    - 1,040,809,705 web pages, in 10 languages
    - 5 TB, compressed. (25 TB, uncompressed.)
  - No way to do this exhaustively
  - Look only at the set of returned documents
    - **Assumption**: if there are enough systems being tested and not one of them returned a document – the document is not relevant

Relevance assessments - Pooling

- Start with result lists retrieved by multiple systems (runs)
- Combine the results retrieved by all systems
- Choose a parameter *k* (typically 100)
- Choose the top *k* documents as ranked in each submitted run
- The pool is the union of these sets of docs
  - Between *k* and (# submitted runs) × *k* documents in pool
  - $(k+1)^{st}$ document returned in one run either irrelevant or ranked higher in another run
- Give pool to judges for relevance assessments



From Donna Harman

Relevance assessments - Pooling

- Conditions under which pooling works [Robertson]
  - Range of different kinds of systems, including manual systems
  - Reasonably deep pools (100+ from each system)
    - But depends on collection size
  - The collections cannot be too big
    - Big is so relative…

Relevance assessments - Pooling

- Advantage of pooling:
  - Fewer documents must be manually assessed for relevance
- Disadvantages of pooling:
  - Can’t be certain that all documents satisfying the query are found (recall values may not be accurate)
  - Runs that did not participate in the pooling may be disadvantaged
  - If only one run finds certain relevant documents, but ranked lower than 100, it will not get credit for these.

Relevance assessments

- Pooling with randomized sampling
- As the data collection grows, the top 100 may not be representative of the entire result set
  - (i.e. the assumption that everything after is not relevant does not hold anymore)
- Add to the pool a set of documents randomly sampled from the entire retrieved set
  - If the sampling is uniform → easy to reason about, but may be too sparse as the collection grows
  - Stratified sampling: get more from the top of the ranked list

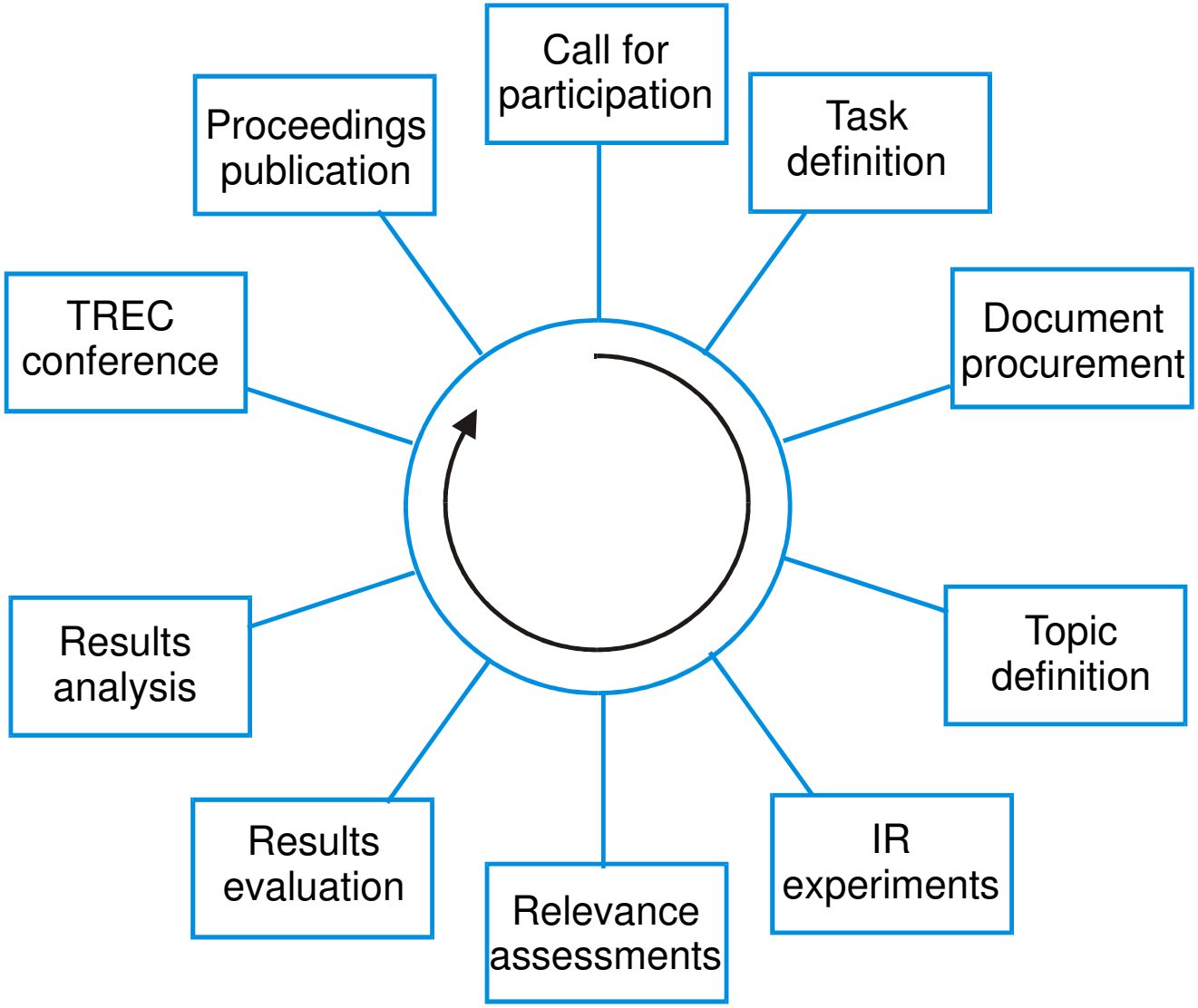
Outline

- Introduction
- Retrieval effectiveness evaluation
- Evaluation campaigns
- User-based evaluation
- Conclusion

How does a typical evaluation campaign run?

- Take ImageCLEF as a "typical" evaluation campaign
- http://www.imageclef.org
- In 2008, it consisted of 5 tasks:
  - photographic retrieval task,
  - medical retrieval task,
  - general photographic concept detection task,
  - medical automatic image annotation task, and
  - image retrieval from a collection of Wikipedia articles.
- We will look at the ImageCLEF 2007 photographic retrieval task
- Opportunity to combine text and visual retrieval algorithms

Circle of events



1. Call for Participation

- Flyer
- Web

UNIVERSITÉ DE GENÈVE

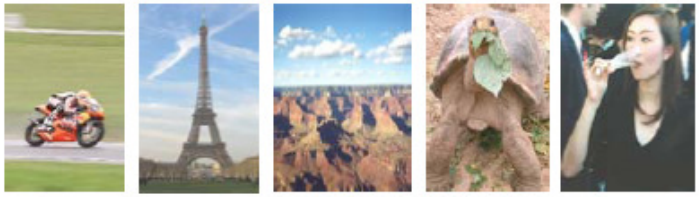
ImageCLEF 2007

Text and/or Content-Based Cross Language Image Retrieval

First Announcement and Call for Participation

Photographic retrieval task

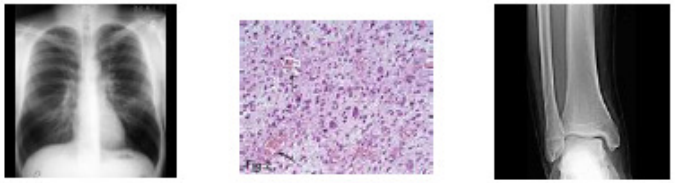
Goal: given a multilingual statement describing a user information need, find as many relevant images as possible from an image collection. This task simulates text-based retrieval from photographs with multilingual captions. Queries for content-based image retrieval will be offered, too.
Image analysis: not required, but can augment text-based retrieval methods and results of an example visual retrieval system will be made available. Visual-only queries will also be provided.
Queries: 50 information needs, each described by a short text in a range of languages including English, Italian, Spanish, French, German, Chinese, Japanese and Russian, and sample images. Several topics will be offered to emphasise both semantic and visual queries.
Collection: 20,000 colour photographs with semi-structured captions in English, German and Spanish.



Challenges: multilingual queries, short caption texts, semi-structured captions in English/German/Spanish,

Medical retrieval task

Goal: given an information need described by medical images and a short text, find other images from the dataset that fulfil this need. The task simulates medical practitioners searching cases similar to one on which they are working; this can be important for evidence-based medicine as well as for teaching .
Image analysis: not required for all tasks, results of an example visual retrieval system will be made available.
Queries: 30 information needs described by a short text and image(s) (visual, mixed and semantic queries).



Collection: ~80,000 medical images from five collections are combined to create a large, heterogeneous resource (English/French/German).
Challenges: combining text and visual methods for retrieval, domain-specific medical terminology and notes of varying quality in mixed target languages.
Aims: to compare methods of visual and text-based retrieval and their complementary influence, to investigate exploitation of heterogeneous annotations, to compare translation methods, retrieval models, and

2. Task definition

- This new challenge allows for the investigation of the following research questions:
  - Are traditional text retrieval methods still applicable for such short captions?
  - How significant is the choice of the retrieval language?
  - How does the retrieval performance compare to retrieval from collections containing fully annotated images (ImageCLEFphoto 2006)?
  - Has the general retrieval performance improved in comparison with retrieval from lightly annotated images (ImageCLEFphoto 2006)?

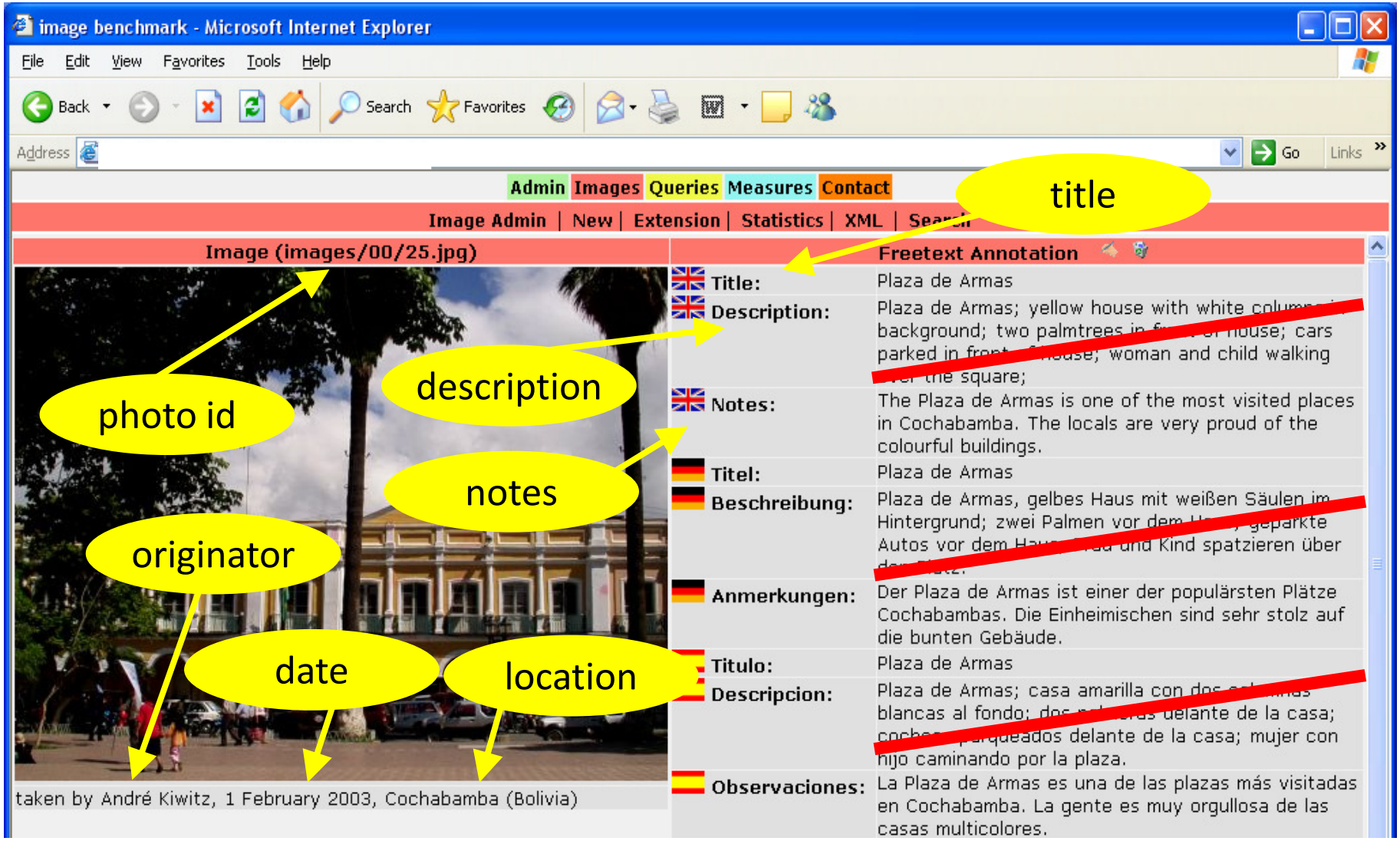
3. Data procurement

- Get a database of images suitable for the task
- Considerations:
  - Copyright restrictions
  - Size
  - Quality
  - Annotations
  - Are realistic queries possible?
- Data drives what can actually be evaluated
- Cannot use the same dataset for too long

- In ImageCLEF 2006–2008, used the IAPR TC-12 image dataset
- Images provided by *Viventura*, a travel company
- Travel guides take photos on tours and upload them to the company website
- Each image annotated in English, German and Spanish by Michael Grubinger

- In 2007, the *description* field was omitted to make it more difficult to search purely using text

IAPR-TC12 example image



4. Topic/Query definition

- Want to have realistic queries/topics
- Type of queries limited by the database used
- More topics lead to more confidence in the experimental results – 50 topics is a commonly used number

- A query log file of a search engine is often a good source of realistic queries

ImageCLEFphoto query topic background

TOPIC BACKGROUND

The following background information has been double-checked with the employees, guides and customers of viventura in order to further back-up the realistic nature of the query topics.

ID	TITLE	LOGFILE	BACKGROUND
1	accommodation with swimming pool	YES	tourists wants to stay only at accommodation with a swimming pool
2	church with more than two towers	YES	tourist did not remember the name of the basilica in quito, but did remember that it has more than two towers
3	religious statue in the foreground	YES	Tourists always take pictures of statues, and they always look for these statues then too. There are many images of churches with many statues on them in the background, so at least one person once typed in "statues in the foreground", which gives us the nice chance to investigate if retrieval systems actually handle the "in the foreground" information well (the "religious" was added to narrow the concept)
4	group standing in front of mountain landscape in Patagonia	YES	Patagonia is one of the most breath-taking regions in the world, and viventura brings their customers to some spots with a very picturesque background, like the Cerro Camapanario in Bariloche (Argentina) or the Torres del Paine National Park, which offer perfect spots for a group photo due to their very scenic backgrounds. It is a very difficult topic though for the systems as it is not trivial for them to find out what locations actually lie in Patagonia, as even the narrative descriptions mention the regions and not the specific location
5	animal swimming	YES	Many tours of viventura include parts in which swimming animals can be seen, especially the trips to the Isla de la Plata during the mating period of the humpback whales, or the trips to the last paradise on earth, the Galapagos Islands. Most users actually query for the specific animal (humpbackwhale swimming), some were general - we used the general version for two reasons: 1) more relevant images for the
6	straight road in the USA	DER	tourist enquired for a group photo taken on the Pan-American Highway in South-America --> this was changed to USA because 1) there are many images of straight roads in the USA in the database, and 2) in order to test the systems' ability to deal with abbreviations.
7	group standing in salt pan	YES	user looking for a group photo in Uyuni, Bolivia, but couldn't remember the name of that specific salt lake
8	host families posing for a photo	YES	language students often stay with host families in order to practice their language skills also outside the classroom. and they, of course, want to see who they will be staying with
9	tourist accomodation near Lake Titicaca	YES	one hotel operator at Puno decided to renovate a couple of rooms although they were already booked an confirmed. the guide had to look for alternative arrangements near Lake Titicaca
10	destinations in Venezuela	YES	a Venezuela tour was organized in 2005 for the first time, and many customers didn't actually know what there is to see in Venezuela
11	black and white photos of Russia	NO	created as a visual topic
12	people observing football match	NO	created to test photos with actions and the discrimating power of query terms (because there are many photos of football matches without spectators). Further, it includes a bit of pettifoggery as relevant images are just images of football (soccer) and no other codes. There are not many countries left which call this sport "soccer", even Australia has officially changed the name to "football". Should people see a
13	exterior view of school building	YES	On the viYoung Peru-Bolivia-Chile tour, the participants visit a school in the Arequipanian suburb of Villa Cerrillos, which was built by viventura as part of one of their social programs. It is a quite colourful, blue and white building with red doors which stands in the middle of light brown desert sand and thus an object that tourists and guides take photos of. Only that the name of the suburb (Villa Cerrillos) gets
14	scenes of footballers in action	NO	see 12
15	night shots of cathedrals	YES	tours visit, for example, the Plaza de Armas in Lima or the Cerro Santa Ana in Guayaquil at night, and tourists normally take pictures of the illuminated cathedrals there.

Queries

- Offered in 16 languages (particular to CLEF):

ID	Topic Title	ID	Topic Title
1	accommodation with swimming pool	31	volcanos around Quito
2	church with more than two towers	32	photos of female guides
3	religious statue in the foreground	33	people on surfboards
4	group standing in front of mountain landscape in Patagonia	34	group pictures on a beach
5	animal swimming	35	bird flying
6	straight road in the USA	36	photos with Machu Picchu in the background
7	group standing in salt pan	37	sights along the Inca-Trail
8	host families posing for a photo	38	Machu Picchu and Huayna Picchu in bad weather
9	tourist accommodation near Lake Titicaca	39	people in bad weather
10	destinations in Venezuela	40	tourist destinations in bad weather
11	black and white photos of Russia	41	winter landscape in South America
12	people observing football match	42	pictures taken on Ayers Rock
13	exterior view of school building	43	sunset over water
14	scenes of footballers in action	44	mountains on mainland Australia
15	night shots of cathedrals	45	South American meat dishes
16	people in San Francisco	46	Asian women and/or girls
17	lighthouses at the sea	47	photos of heavy traffic in Asia
18	sport stadium outside Australia	48	vehicle in South Korea
19	exterior view of sport stadia	49	images of typical Australian animals
20	close-up photograph of an animal	50	indoor photos of churches or cathedrals
21	accommodation provided by host families	51	photos of goddaughters from Brazil
22	tennis player during rally	52	sports people with prizes
23	sport photos from California	53	views of walls with asymmetric stones
24	snowcapped buildings in Europe	54	famous television (and telecommunication) towers
25	people with a flag	55	drawings in Peruvian deserts
26	godson with baseball cap	56	photos of oxidised vehicles
27	motorcyclists racing at the Australian Motorcycle Grand Prix	57	photos of radio telescopes
28	cathedrals in Ecuador	58	seals near water
29	views of Sydney's world-famous landmarks	59	creative group pictures in Uyuni
30	room with more than two beds	60	salt heaps in salt pan

Typical query example

```
<top>
<num> Number: 4 </num>
<title> group in front of mountain landscape </title>
<narr> Relevant images will show a group of at least
three people in front of a mountain landscape. Images
with a single person or a couple are not relevant,
and images that do not show at least two mountains in
the background are not relevant either. </narr>
<image> images/03/3474.jpg </image>
<image> images/09/9882.jpg </image>
<image> images/23/23221.jpg </image>
</top>
```

Query

Narrative

Example images



- Narrative aims to help with relevance judgements, is however often useful for search too

5. IR experiments

- Make the dataset and queries available to the participants
- Make available a result file format
  - For each query, the participant returns a ranked list of the images that best respond to the query
- Each participant may submit a number of runs (attempts with different parameters)
  - Each run has the retrieval results for all queries
- Set a deadline for submissions

6. Relevance assessments

- We need to know the ground truth for every query
  - Which documents are relevant and which not?
  - Most often a binary value – relevant/not relevant – is used, but degrees of relevance are also possible

- Obvious solution is to manually assess every document for every query
- With 20000 images, 50 queries, ±5 seconds per image: need 58 days (working 24 hours per day)
- Pooling is used to speed up relevance assessments

7. Results evaluation

- Evaluation measures calculated for every submitted run.
  - Precision
  - Recall
  - Mean Average Precision (MAP)

Results for imageCLEFphoto 2007

- 20 participants

Best result for each query and caption language combination

Query (Caption)	Group/Run ID	MAP
English (English)	CUT/cut-EN2EN-F50	0.318
German (English)	XRCE/DE-EN-AUTO-FB-TXTIMG_MPRF	0.290
Portuguese (English)	Taiwan/NTU-PT-EN-AUTO-FBQE-TXTIMG	0.282
Spanish (English)	Taiwan/NTU-ES-EN-AUTO-FBQE-TXTIMG	0.279
Russian (English)	Taiwan/NTU-RU-EN-AUTO-FBQE-TXTIMG	0.273
Italian (English)	Taiwan/NTU-IT-EN-AUTO-FBQE-TXTIMG	0.271
S. Chinese (English)	CUT/cut-ZHS2EN-F20	0.269
French (English)	Taiwan/NTU-FR-EN-AUTO-FBQE-TXTIMG	0.267
T. Chinese (English)	Taiwan/NTU-ZHT-EN-AUTO-FBQE-TXTIMG	0.257
Japanese (English)	Taiwan/NTU-JA-EN-AUTO-FBQE-TXTIMG	0.255
Dutch (English)	INAOE/INAOE-NL-EN-NaiveWBQE-IMFB	0.199
Swedish (English)	INAOE/INAOE-SV-EN-NaiveWBQE-IMFB	0.199
Visual (English)	INAOE/INAOE-VISUAL-EN-AN_EXP_3	0.193
Norwegian (English)	DCU/NO-EN-Mix-sgramRF-dyn-equal-fire	0.165
German (German)	Taiwan/NTU-DE-DE-AUTO-FBQE-TXTIMG	0.245
English (German)	XRCE/EN-DE-AUTO-FB-TXTIMG_MPRF_FLR	0.278
Swedish (German)	DCU/SW-DE-Mix-dictRF-dyn-equal-fire	0.179
Danish (German)	DCU/DA-DE-Mix-dictRF-dyn-equal-fire	0.173
French (German)	CUT/cut-FR2DE-F20	0.164
Norwegian (German)	DCU/NO-DE-Mix-dictRF-dyn-equal-fire	0.167
Spanish (Spanish)	Taiwan/NTU-ES-ES-AUTO-FBQE-TXTIMG	0.279
English (Spanish)	CUT/cut-EN2ES-F20	0.277
German (Spanish)	Berkeley/Berk-DE-ES-AUTO-FB-TXT	0.091
English (Random)	DCU/EN-RND-Mix-sgramRF-dyn-equal-fire	0.168
German (Random)	DCU/DE-RND-Mix-sgram-dyn-equal-fire	0.157
French (Random)	DCU/FR-RND-Mix-sgram-dyn-equal-fire	0.141
Spanish (Random)	INAOE/INAOE-ES-RND-NaiveQE-IMFB	0.124
Dutch (Random)	INAOE/INAOE-NL-RND-NaiveQE	0.083
Italian (Random)	INAOE/INAOE-IT-RND-NaiveQE	0.080
Russian (Random)	INAOE/INAOE-RU-RND-NaiveQE	0.076
Portuguese (Random)	INAOE/INAOE-PT-RND-NaiveQE	0.030
Visual	XRCE/AUTO-NOFB-IMG_COMBFK	0.189

Average results by retrieval modality

Modality	MAP
Mixed	0.149 (0.066)
Text Only	0.120 (0.040)
Image Only	0.068 (0.039)

8. Results analysis

- What can you learn from the results?
- Relate to the research questions in step 2

- Evaluation measures are only comparable if the experiments are carried out
  - On the same test database
  - Using the same set of queries
- Be careful comparing results in different years

9. Conference

- Best results, but also interesting approaches are presented
  - Possible dilemma for organisers: What if the best system is the same as in the previous year?
- Discussion between participants

10. Proceedings

Overview of the ImageCLEFphoto 2007 photographic retrieval task

Michael Grubinger[1], Paul Clough[2], Allan Hanbury[3], Henning Müller[4]

[1] Victoria University, Melbourne, Australia
[2] Sheffield University, Sheffield, UK
[3] Vienna University of Technology, Vienna, Austria
[4] University and Hospitals of Geneva, Switzerland

Abstract

ImageCLEFphoto 2007 is the general photographic ad-hoc retrieval task of the *ImageCLEF 2007* evaluation campaign and provides both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information retrieval from generic photographic collections. In 2007, the evaluation objective concentrated on retrieval of lightly annotated images, a new challenge that attracted a large number of submissions: a total of 20 participating groups submitting a record number of 616 system runs. This paper summarises the components used in the benchmark, including the document collection, the search tasks, an analysis of the submissions from participating groups, and results.

Outline

- Introduction
- Retrieval effectiveness evaluation
- Evaluation campaigns
- User-based evaluation
- Conclusion

Laboratory experiments

- Abstraction from the real world in well controlled laboratory conditions
- Goal is retrieval of items of information
- Rigorous testing
- Over-constrained
- Can obtain scientifically reliable results

- But how does this relate to the real world?
  - Information needs are often related
  - Workflow
  - …

- Comparison to a tennis racket:
  - No evaluation of the device will tell you how well it will perform in real life – that largely depends on the user
  - But the user will chose the device based on the lab evaluation

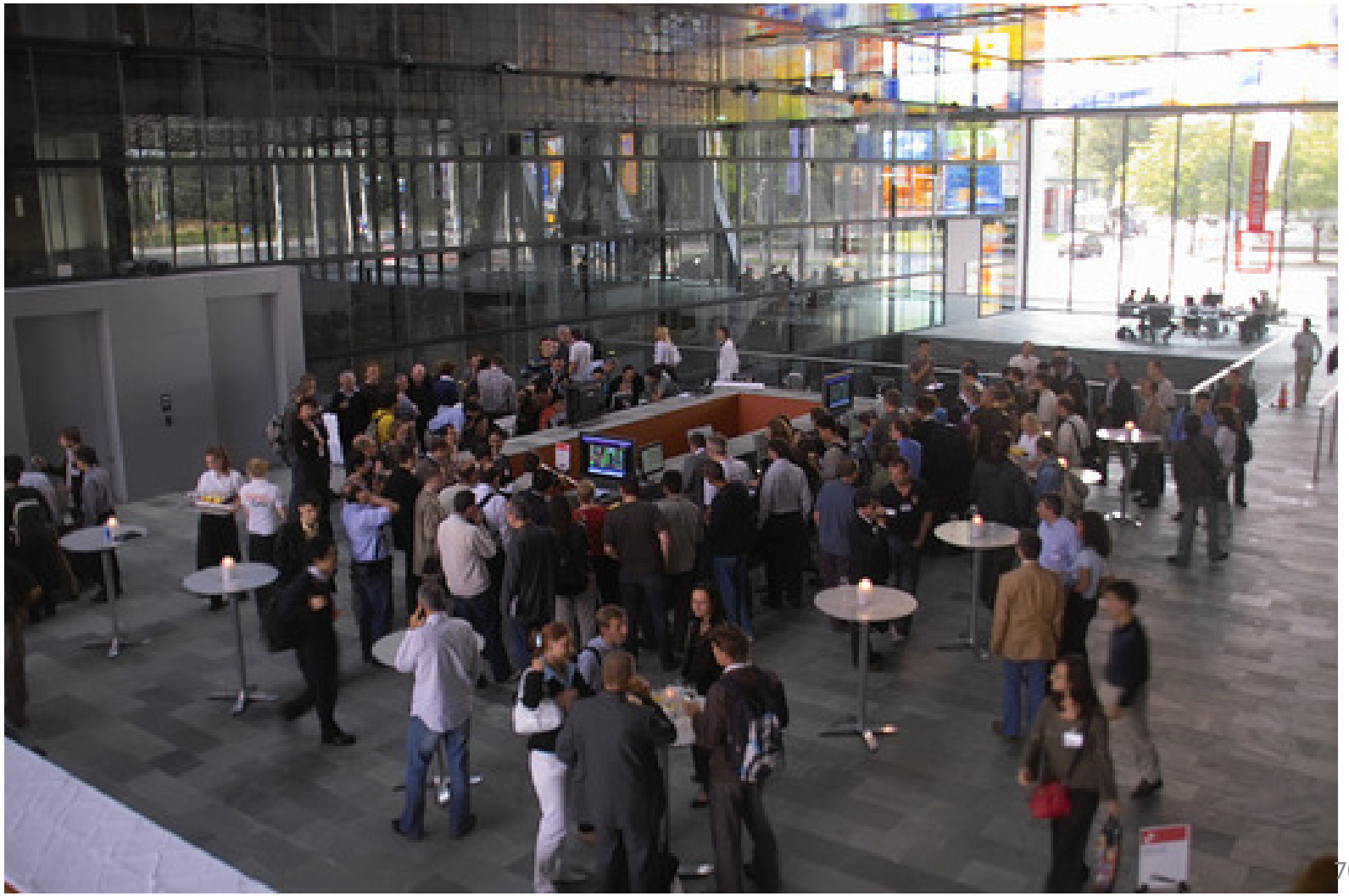
User-based evaluation

- 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
     - Measures the user's satisfaction with the system
- Some work on 1., very little on 2.

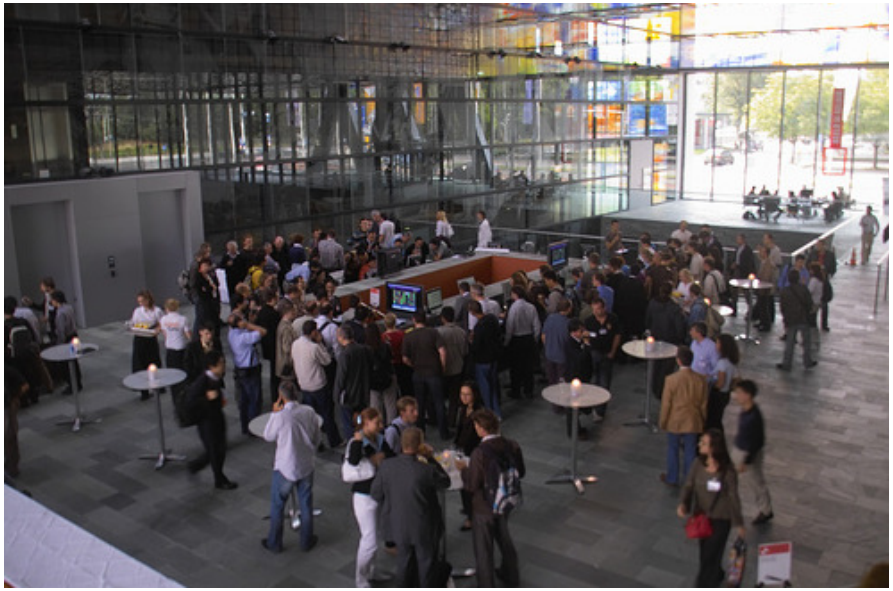
User satisfaction evaluation

- Expensive and difficult to do correctly.
  - Large, representative sample of actual users
  - Each system must be equally well developed and have a user interface
  - Each participant must be equally well trained on each system
  - The learning effect must be controlled for
- User satisfaction is very subjective
  - UIs play a major role
  - Search dissatisfaction can be a result of the non-existence of relevant documents

Beyond the Laboratory: VideOlympics



VideOlympics 2007



- TrecVID 2005/2006 data
- Example queries:

Find shots of one or more helicopters in flight.

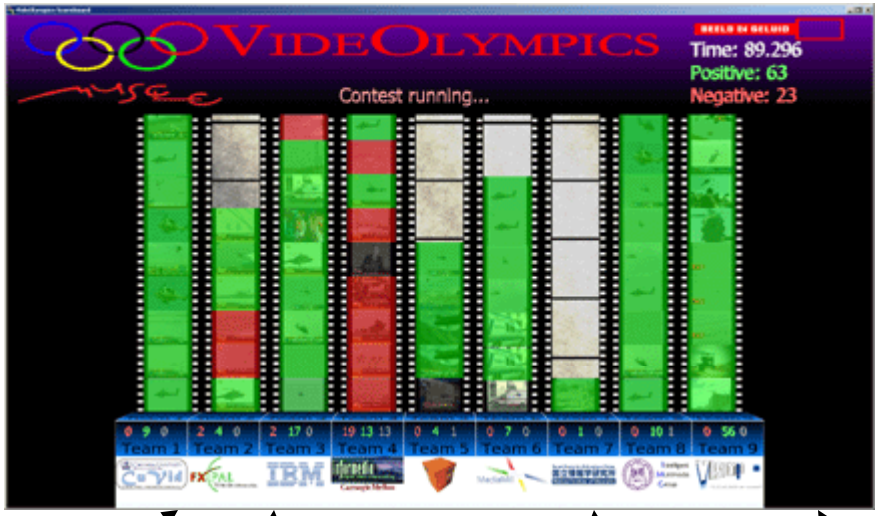
Find shots of a hockey rink with at least one of the nets fully visible from some point of view.

Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

Find shots of a group including at least four people dressed in suits, seated, and with at least one flag.

VideOlympics Result Display System

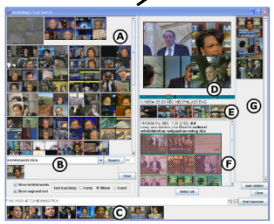
One display

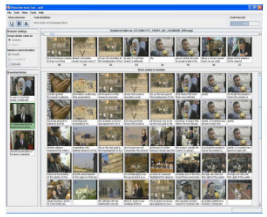


High penalty for wrong results

Run ends when 100 results have been found or 5 minutes have past

TRECVID like queries
A result is submitted as soon as it is found





……

Retrieval Systems running on Notebooks

Conclusion

- IR Evaluation is a research field in itself
- Without evaluation, research is pointless
- Most IR Evaluation exercises are laboratory experiments
  - As such, care must be taken to match, to the extent possible, real needs of the users

Bibliography

- Modern Information Retrieval
  - R. Baeza-Yates, B. Ribeiro-Neto
- TREC – Experiment and Evaluation in Information Retrieval
  - E. Voorhees, D. Harman (eds.)
- A Comparison of Statistical Significance Tests for Information Retrieval Evaluation
  - M. Smucker, J. Allan, B. Carterette (CIKM'07)
- A Simple and Efficient Sampling Method for Estimating AP and NDCG
  - E. Yilmaz, E. Kanoulas, J. Aslam (SIGIR'08)

Bibliography

- *Do User Preferences and Evaluation Measures Line Up?*, M. Sanderson and M. L. Paramita and P. Clough and E. Kanoulas 2010
- *A Review of Factors Influencing User Satisfaction in Information Retrieval*, A. Al-Maskari and M. Sanderson 2010
- *Towards higher quality health search results: Automated quality rating of depression websites*, D. Hawking and T. Tang and R. Sankaranarayana and K. Griffiths and N. Craswell and P. Bailey 2007
- *Evaluating Sampling Methods for Uncooperative Collections*, P. Thomas and D. Hawking 2007
- *Comparing the Sensitivity of Information Retrieval Metrics*, F. Radlinski and N. Craswell 2010
- *Redundancy, Diversity and Interdependent Document Relevance*, F. Radlinski and P. Bennett and B. Carterette and T. Joachims 2009
- *Does Brandname influence perceived search result quality? Yahoo!, Google, and WebKumara*, P. Bailey and P. Thomas and D. Hawking 2007
- *Methods for Evaluating Interactive Information Retrieval Systems with Users*, D. Kelly 2009
- *C-TEST: Supporting Novelty and Diversity in TestFiles for Search Tuning*, D. Hawking and T. Rowlands and P. Thomas 2009
- *Live Web Search Experiments for the Rest of Us*, T. Jones and D. Hawking and R. Sankaranarayana 2010
- *Quality and relevance of domain-specific search: A case study in mental health*, T. Tang and N. Craswell and D. Hawking and K. Griffiths and H. Christensen 2006
- *New methods for creating testfiles: Tuning enterprise search with C-TEST*, D. Hawking and P. Thomas and T. Gedeon and T. Jones and T. Rowlands 2006
- *Test Collection Based Evaluation of Information Retrieval Systems*, M. Sanderson 2010