



University of Twente

Information Retrieval Modeling

Russian Summer School in Information Retrieval

Djoerd Hiemstra

<http://www.cs.utwente.nl/~hiemstra>



University of Twente
The Netherlands



PART 5

Evaluation

Course material

- Djoerd Hiemstra and Wessel Kraaij, Evaluation of Multimedia Retrieval Systems, In Multimedia Retrieval, Springer, pages 347-365, 2007

Goal

- An introduction to doing *real* (measurable, repeatable) research
- Getting acquainted with the TREC paradigm

Science?

- At some point during your studies (or in your professional life), you might have to do a scientific evaluation...
 - choose data
 - choose suitable evaluation measures: assign values to results of your system
 - choose a statistical methodology: determine whether observed differences are significant

The traditional IR experiment

- To start with you need
 - An IR system (or two)
 - A collection of documents
 - A collection of requests
- Then you run your experiment
 - Input (index) the documents
 - Put each request to the system
 - Collect the output

The traditional IR experiment

- Then you need to
 - Evaluate the output, document by document
 - Discover (??) the good documents your system has missed
 - Analyse the results

The traditional IR experiment

- What is a document?
 - traditionally: a package of information structured by an author
- What is a request?
 - a description of a topic of interest
 - a partial representation of an underlying information need
- What is a system?
 - a device that accepts a request and delivers or identifies documents
 - "device" may be an organisation: involve people(!)

The traditional IR experiment

- Assuming that documents are either relevant or not, the objective is:
 - To retrieve relevant documents
 - Not to retrieve non-relevant documents


[Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#) [more »](#)

black jaguar

Search

[Advanced Image Search](#)
[Preferences](#)

Moderate SafeSearch is on

Images

Results 1 - 20 of about 5,660 for [black jaguar](#). (0.10 seconds)Show: [All sizes](#) - [Large](#) - [Medium](#) - [Small](#)

Jaguar-XK180-01_800.jpg
 800 x 600 pixels - 49k
www.auto-sfondi-desktop.com/Wallpapers_Jaguar...



jaguar-xk8-2-600.jpg
 600 x 450 pixels - 132k
www.dtconcoours.com/.../jaguar-xk8-2-600.jpg



animal_jaguar_jump.jpg
 226 x 225 pixels - 12k
www.sandiegozoo.org/kids/animal_profile_jagua...



Black Jaguar full.jpg
 322 x 450 pixels - 31k
www.jrsfilm.com/blackJaguar.asp



10900843yzxtRkFVaW_ph.jpg
 504 x 468 pixels - 26k
community.webshots.com/.../10900843yzxtRkFVaW



belize - black jaguar.jpg
 1024 x 768 pixels - 138k
www.stanford.edu/~wgupta/images/belize%20-%20...



twilight00.jpg
 504 x 316 pixels - 30k
www.cathouse-fcc.org/images/twilight00.jpg
[\[More results from www.cathouse-fcc.org \]](#)



jaguar_black.gif
 480 x 328 pixels - 122k
www.mongabay.com/.../belizejaguar_black.html



vw-Black-Jaguar.jpg
 450 x 251 pixels - 39k
www.linkandpinhobbies.com/Graphics/vw-Black-J...



black-jaguar.jpg
 292 x 194 pixels - 12k
www.ecology.info/ecology-jaguar-puma.htm



animaSDZ_0150-BlackJaguar-Pant...
 208 x 312 pixels - 17k
animals.timduru.org/dirlist/panther/SDZ_0150-...



S5300012.jpg
 640 x 480 pixels - 63k
photos1.blogger.com/img/190/1614/640/S5300012.jpg



10900843yzxtRkFVaW_ph.jpg
 504 x 468 pixels - 26k
community.webshots.com/.../10900843yzxtRkFVaW



Black_Jaguar_copy.jpg
 575 x 457 pixels - 26k
www.jesuspaintings.com/wildlife_paintings.htm



black jaguar.jpg
 600 x 450 pixels - 80k
biology.kenyon.edu/.../Part%203%20Vehicles.htm

igs.htm

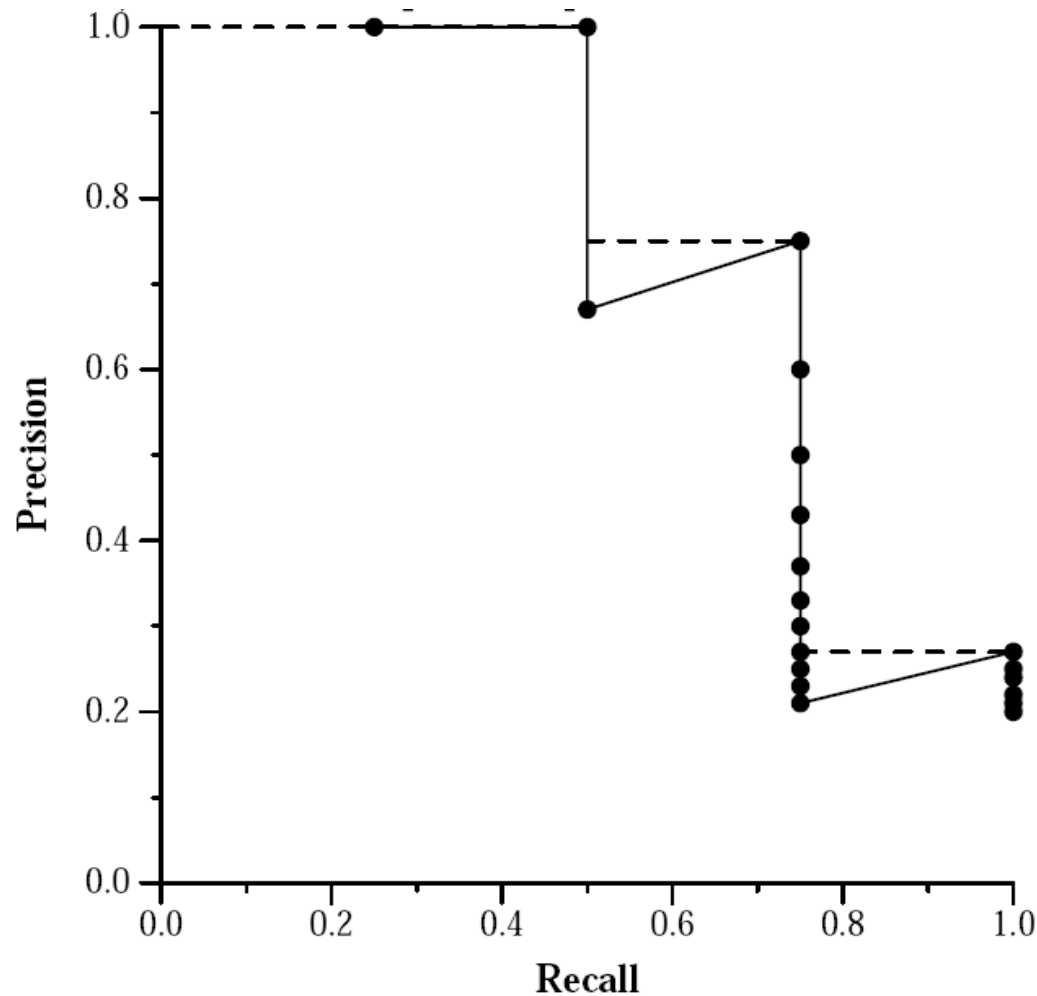
The traditional IR experiment

- Evaluation measures
 - precision = r/n : fraction of retrieved documents that is relevant
 - recall = r/R : fraction of relevant documents that is retrieved
- r : number of relevant documents retrieved
 n : number of documents retrieved
 R : number of relevant documents

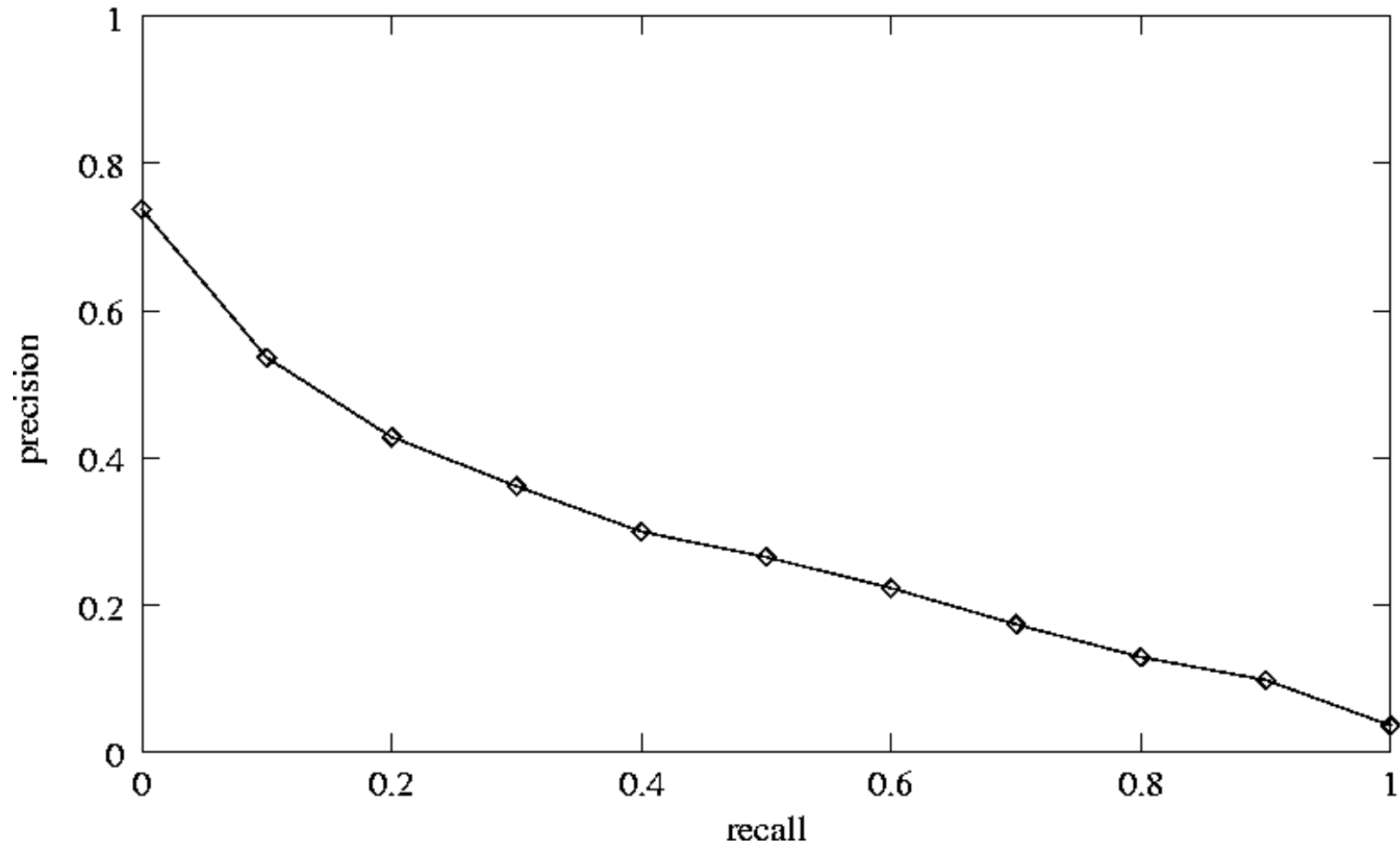
What about ranked output?

- Report precision for positions in the ranked list
 - 5, 10, 20 document retrieved
- Report precision for some recall levels
 - precision at 0.1, 0.2, etc.

Recall-precision plot



Recall-precision plot



The traditional IR experiment

- Problems with IR system evaluation
 - costly (involves users)
 - which documents did the system miss?
 - hard to repeat in same settings (learning / fatigue effects)
 - we need a complete system(!) we do not in general know how to evaluate components



The TREC paradigm

doing laboratory tests

Benchmark collections

- Consists of three parts:
 - documents (realistic contents and size)
 - requests (textual description of information need; realistic, "real" application)
 - relevance assessments: how useful is the retrieved document?
- How to design?
 - Cranfield → TREC → CLEF, INEX, ROMIP

What is TREC?

- Competition/collaboration between IR research groups world-wide
- Run by the US National Standards and Technology
- TREC provides:
 - common test collections
 - common tasks
 - common measures
 - common evaluation procedure

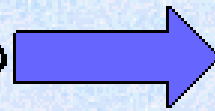


What is TREC?

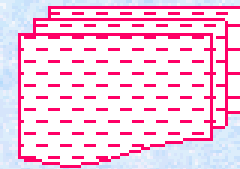
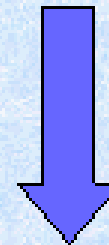
- A workshop series that provides the infrastructure for large-scale testing of text retrieval technology
 - realistic test collections
 - uniform, appropriate scoring procedures
 - a forum for the exchange of research ideas and for the discussion of research methodology

TREC approach

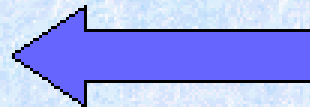
Assessors create topics at NIST



Topics are sent to participants, who return ranking of best 1000 documents per topic



Systems are evaluated using relevance judgments



NIST forms pools of unique documents from all submissions which the assessors judge for relevance



An example TREC topic

<top>

<num> 405

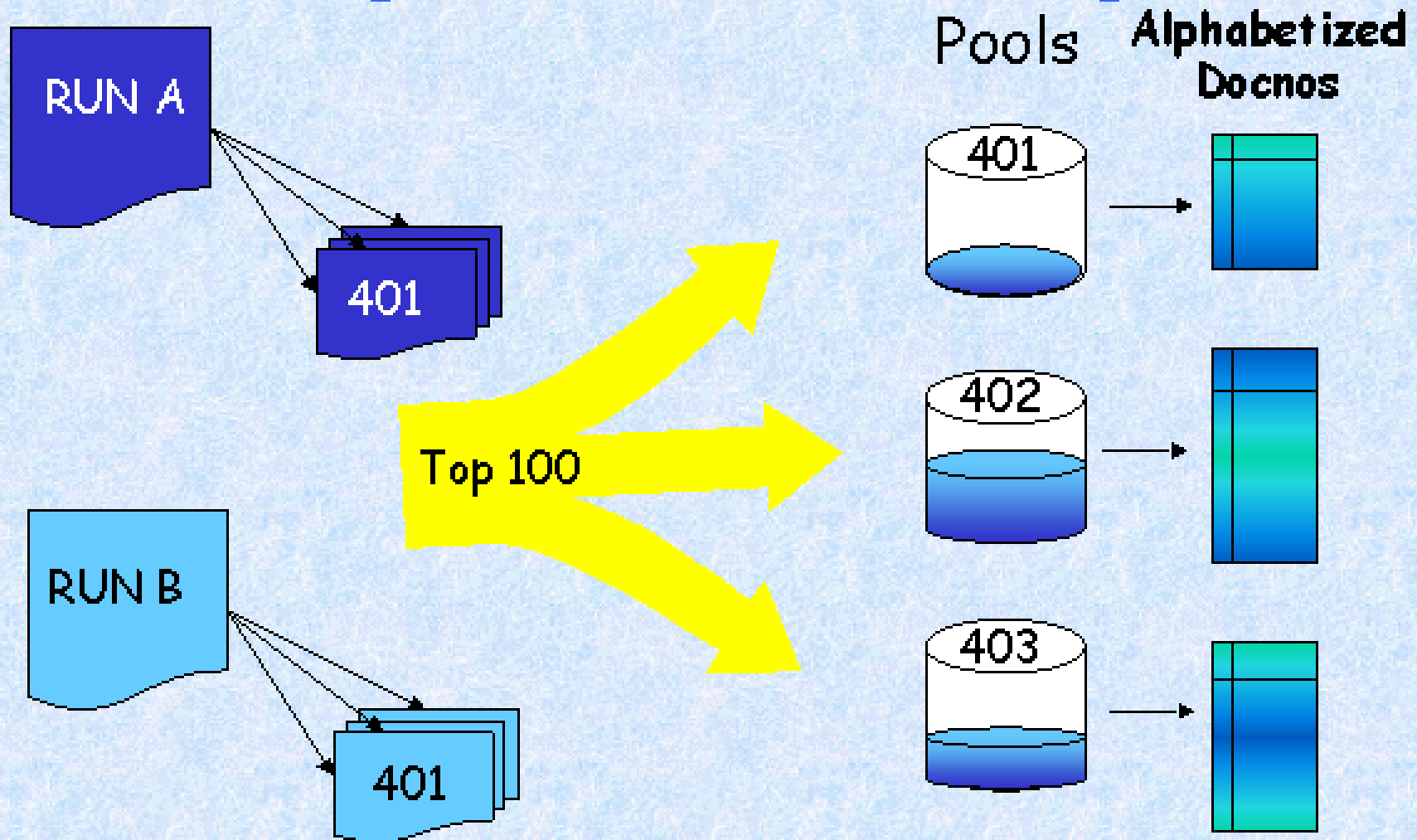
<title> cosmic events

<desc> What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?

<narr> New theories or new interpretations concerning known celestial objects made as a result of new technology are not relevant.

</top>

Creating Relevance Judgments





(thanks to Ellen Voorhees)

Text REtrieval Conference (TREC)



TREC assumptions about relevance

- Relevance of one element does not affect the relevance of another element
- Relevance is a binary decision, i.e., a document is either relevant or not
- A document is relevant if it would help in writing an article about the subject
 - relevant? topicality? clarity? recency? accuracy? trustworthiness?

TREC assumptions about systems

- A system is a programme
 - the user is outside the system
- A system is an input-output device
 - query in, documents out
 - most real searches involve interaction

Quality of a test collection?

- Two concerns:
 - Consistency of the judgements: *do the results of the experiments critically depend on the particular choices of human judges?*
 - Completeness of the judgements: *do the results critically depend on the pool construction process, i.e. on the systems that participated in INEX?*

Consistency of the judgements

- Experiment: 10 topics assessed twice by two different assessors
- Dutch CLEF collection, overlap: 0.465
- TREC: overlap between: 0.421 and 0.494
(Overlap = size of intersection of the relevant document sets divided by the size of the union of the relevant document sets.)
- (Overall agreement 93.4 %)

Completeness of judgements

- Can we use the collection for future experiments?
- What if my run is not judged?
- Experiment: recompute for each official run the average precision as if it was not in the pool, i.e. ignoring the relevant documents uniquely found by that run

Completeness: What if my run is not judged?

run name	unjudged / judged		avg.prec.	difference	unique rel.
ut1	0.4222	0.4230	0.0008	0.2 %	55
aplmonla	0.3943	0.4002	0.0059	1.5 %	29
tnonn3	0.3914	0.3917	0.0003	0.1 %	2
humNL01x	0.3825	0.3831	0.0006	0.2 %	5
tlrnltl	0.3760	0.3775	0.0015	0.4 %	10
tnoen1	0.3246	0.3336	0.0090	2.8 %	32
AmsNIM	0.2770	0.2833	0.0063	2.3 %	32
aplbiennl	0.2692	0.2707	0.0015	0.6 %	7
oce2	0.2363	0.2405	0.0042	1.8 %	21
glaenl	0.2113	0.2123	0.0010	0.5 %	8
oce1	0.2024	0.2066	0.0042	2.1 %	23
medialab	0.1600	0.1640	0.0040	2.5 %	23
EidNL2001A	0.1339	0.1352	<u>0.0013</u>	<u>1.0 %</u>	<u>8</u>
mean:			0.0031	1.2 %	20
standard deviation:			0.0027	1.0 %	15

Significance testing

- When is one system better than another?
 - Maybe the average difference can be contributed to chance?
 - Need a reasonable amount of queries (e.g. 50), which should be a random sample of all possible queries for a given task

Significance testing

- Two hypotheses
 - null-hypothesis H_0 : there is no difference between system A and system B
 - alternative hypothesis H_1 : either system A consistently outperforms system B , or system B consistently outperforms system A
- Show that, given the evaluation results, H_0 is indefensible

Significance testing

- Test statistics should behave differently under H_0 than under H_1 :
 - Paired tests: for each query the performance difference between system A and B consist of a mean difference μ and some error.
 - $H_0 : \mu = 0$; $H_1 : \mu \neq 0$;
 - Paired t-test: assumes that errors are normally distributed. Under H_0 the distribution is Student's t
 - Paired sign test: assumes equal probability of positive and negative error. Under H_0 the distribution is binomial

Conclusion

- To evaluate your system, use a benchmark collection.
- Choose appropriate evaluation measures
- Base your conclusions on statistical tests

References

- Stephen Robertson. Evaluation in information retrieval. Lecture Notes in Computer Science, Springer, pp. 81-92. 2000.
- Ellen Voorhees and Lori Buckland, The Sixteenth Text REtrieval Conference Proceedings (TREC 2007). NIST, 2007
- Donna Harman and/or Ellen Voorhees. Appendix: Common evaluation measures. In “Proceedings of most of the TREC proceedings”.

Acknowledgements

- Thanks to the following people for making their slides available
 - Stephen Robertson (Microsoft Research)
 - Ellen Voorhees (NIST)