Experiment Design and Evaluation for Information Retrieval

Rishiraj Saha Roy | Computer Scientist, Adobe Research Labs India | rroy@adobe.com

Introduction

- Information retrieval (IR) is a *scientific* discipline
- Experimentation is a critical component of the scientific method
- Poor experimental methodologies are not scientifically sound and should be avoided
- IR has become plagued with weak experimentation, causing:
  - Outsiders to think of IR as non-scientific
  - A plethora of minor improvements vs. weak baselines
- All IR experimentation require an *evaluation* of results

Evaluation: Basics

- How do we evaluate the retrieved results?
- “Measure” user happiness
- Happiness is hard to quantify
- Has to be broken down into quantifiable features
- User query: *india commonwealth games 2014*

Google india commonwealth games 2014

Web News Images Videos Maps More Search tools

About 4,39,00,000 results (0.27 seconds)

News for **india commonwealth games 2014**

Watch Live Streaming: **Commonwealth Games 2014** – ...
India.com - 9 minutes ago
Glasgow 2014: Day 8 of the 20th **Commonwealth Games** proved to be quite fruitful for **India** as Discus thrower Vikas Gowda won first gold of ...

Commonwealth Games 2014: **India** Beat Scotland to Finish ...
NDTVSports.com - 5 hours ago

Vikas Gowda Wins **India's** First Athletics Gold at ...
NDTVSports.com (blog) - 11 hours ago

More news for **india commonwealth games 2014**

2014 Commonwealth Games Day 8 Highlights - Sports News
sports.ndtv.com › CWG 2014 › News
15 hours ago - Discus thrower Vikas Gowda won **India's** first gold medal from athletics in **2014 Commonwealth Games**. Catch all the live updates here.

2014 Commonwealth Games Day 9 Live Blog - Sports News
sports.ndtv.com › CWG 2014 › News
by Ashish Maggo - 30 mins ago - With three days remaining in the **2014 Commonwealth Games** in **Glasgow**, **India** will look for a golden home run from boxers, track and field ...

India at the **2014 Commonwealth Games** - Wikipedia, the ...
en.wikipedia.org/wiki/**India**_at_the_**2014_Commonwealth_Games**
The **Glasgow** Games will have 17 sports and 261 medal events and **India** will not be fielding athletes only in three disciplines—netball, rugby sevens and ...
Aquatics - Athletics - Badminton - Boxing

Live Blog - **Glasgow Commonwealth Games 2014**: Day 8 ...
timesofindia.indiatimes.com/.../**commonwealth-games-2014**/.../3934785...
19 hours ago - **India** added three golds to its tally and moved to fifth spot in medals tally on Day 8 of the CWG as wrestlers Yogeshwar Dutt and Babita Kumari ...

India at **Glasgow** - **Commonwealth Games 2014** ...

Relevant

Non-Relevant

Non-Relevant

Relevant

Relevant

Non-Relevant

Happiness Factors: Relevance

Happiness Factors: Layout

Google india commonwealth games 2014

Web News Images Videos Maps More Search tools

About 4,39,00,000 results (0.27 seconds)

News for india commonwealth games 2014

Watch Live Streaming: Commonwealth Games 2014 – ...
India.com - 9 minutes ago
Glasgow 2014: Day 8 of the 20th **Commonwealth Games** proved to be quite fruitful for **India** as Discus thrower Vikas Gowda won first gold of ...

Commonwealth Games 2014: **India** Beat Scotland to Finish ...
NDTVSports.com - 5 hours ago

Vikas Gowda Wins **India's** First Athletics Gold at ...
NDTVSports.com (blog) - 11 hours ago

More news for **india commonwealth games 2014**

2014 Commonwealth Games Day 8 Highlights - Sports News
sports.ndtv.com › CWG 2014 › News
15 hours ago - Discus thrower Vikas Gowda won **India's** first gold medal from athletics in **2014 Commonwealth Games**. Catch all the live updates here.

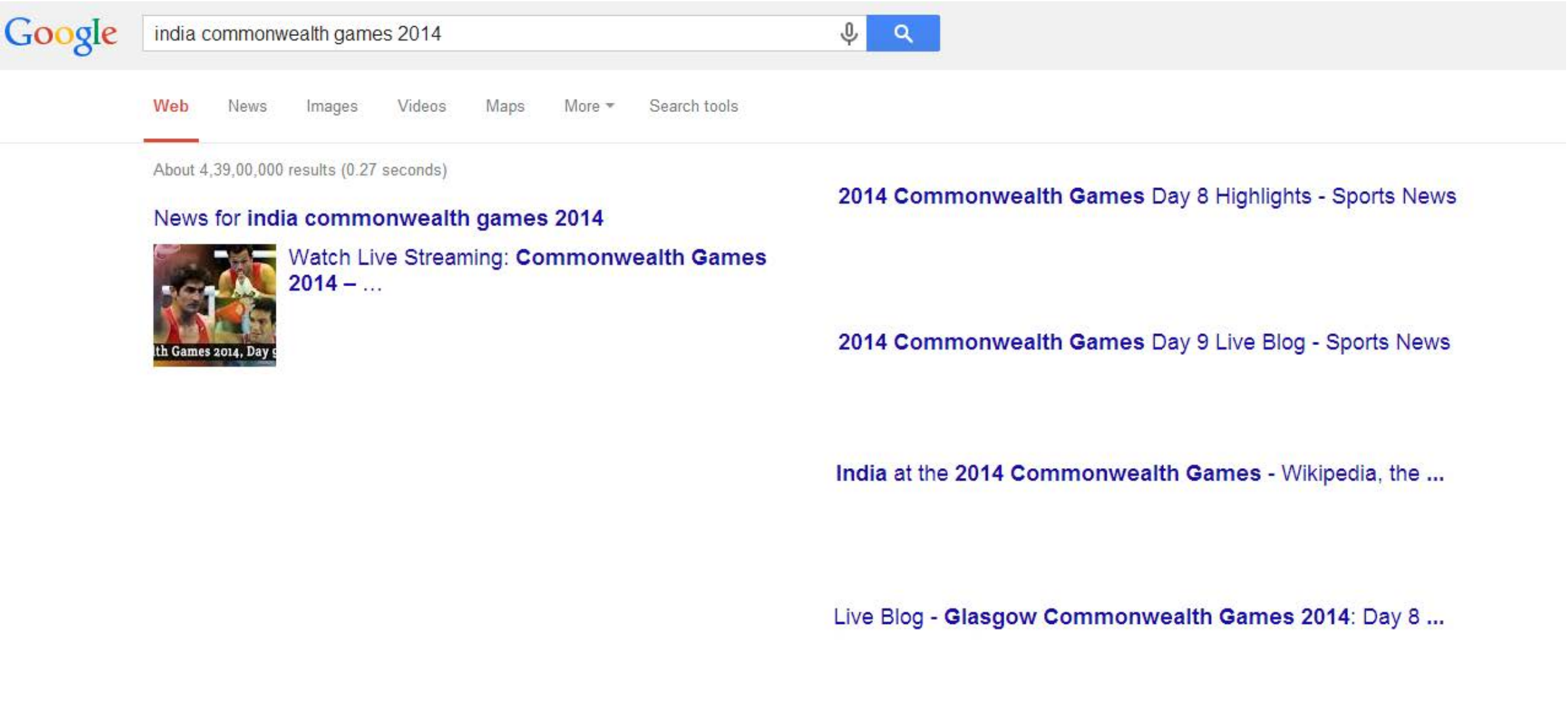
2014 Commonwealth Games Day 9 Live Blog - Sports News
sports.ndtv.com › CWG 2014 › News
by Ashish Maggo - 30 mins ago - With three days remaining in the **2014 Commonwealth Games** in **Glasgow**, **India** will look for a golden home run from boxers, track and field ...

India at the 2014 Commonwealth Games - Wikipedia, the ...
en.wikipedia.org/wiki/**India**_at_the_**2014_Commonwealth_Games**
The **Glasgow** Games will have 17 sports and 261 medal events and **India** will not be fielding athletes only in three disciplines—netball, rugby sevens and ...
Aquatics - Athletics - Badminton - Boxing

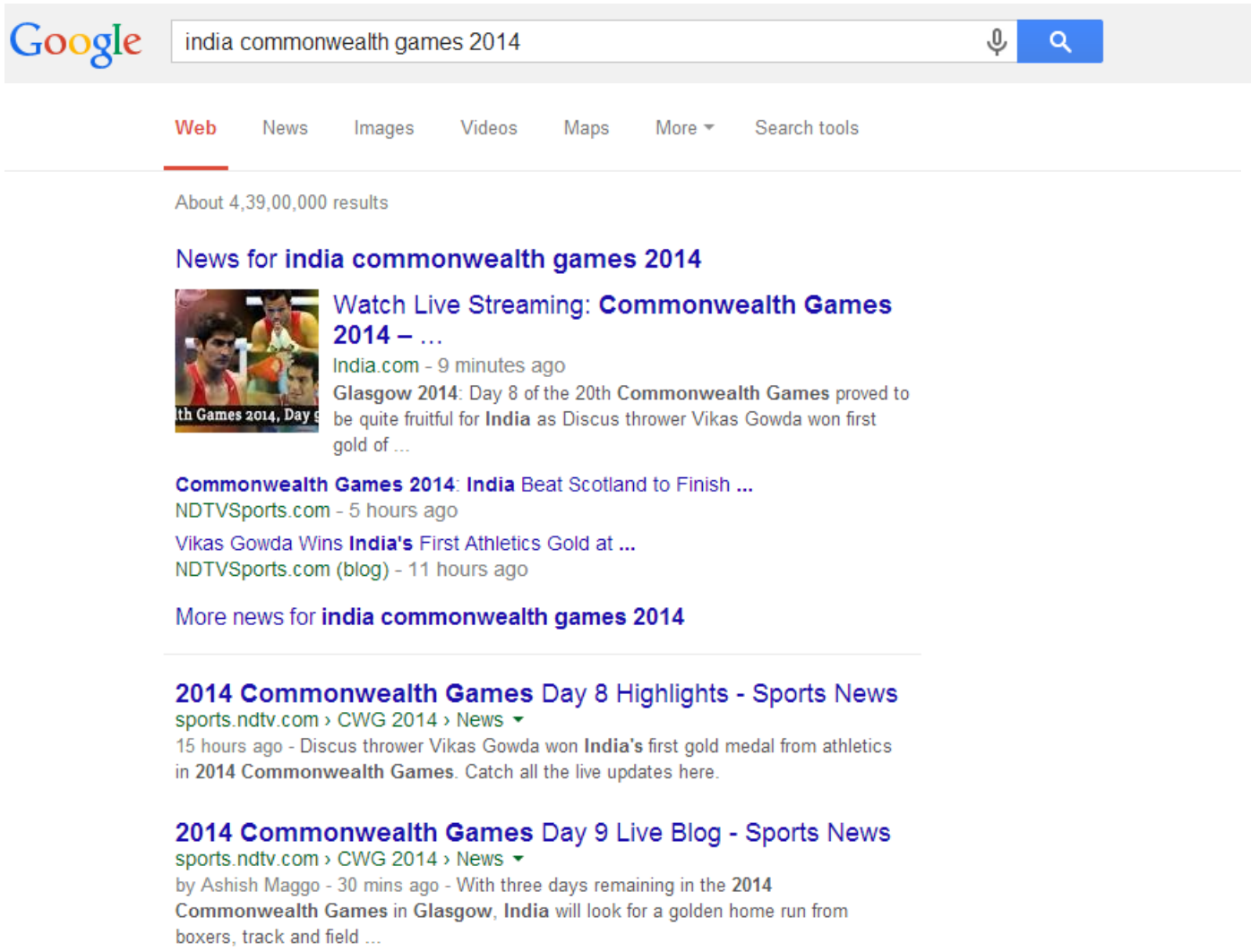
Live Blog - Glasgow Commonwealth Games 2014: Day 8 ...
timesof**india**.**indiatimes**.com/.../**commonwealth-games-2014**/.../3934785...
19 hours ago - **India** added three golds to its tally and moved to fifth spot in medals tally on Day 8 of the CWG as wrestlers Yogeshwar Dutt and Babita Kumari ...

Adobe

Happiness Factors: Content



Happiness Factors: Time



The Cranfield Paradigm

- Our focus: Result relevance
- Relevance measurement requires 3 elements:
  - A benchmark document collection
  - A benchmark suite of queries
  - A (usually) binary assessment of either Relevant or Nonrelevant for each query and each document

Basics

- An **information need** is translated into a **query**
- Relevance is assessed relative to the **information need** and *not* the **query**
- E.g. Information need: *I'm looking for the Windows installer for the Open Office software.*
- Query: ***open office windows***
- Evaluate whether the doc addresses the information need, not whether it has these words

Unordered results: Precision and Recall

- Set-based metrics
- **Precision**: Fraction of retrieved docs that are relevant = P(relevant | retrieved)
- **Recall**: Fraction of relevant docs that are retrieved = P(retrieved | relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Accuracy

- Given a query, an engine classifies each doc as “Relevant” or “Nonrelevant”
- The **accuracy** of the retrieval system is the fraction of these classifications that are correct
  - (tp + tn) / (tp + fp + fn + tn)
- **Accuracy** is a commonly used evaluation measure in classification
- Why is this not a very useful evaluation measure in IR?
- What happens if you return nothing?
- People doing information retrieval *want to find something* and have a some tolerance for non-relevant pages

Precision and Recall

- You can get high recall (but low precision) by retrieving all docs for all queries!
- Recall is a non-decreasing function of the number of docs retrieved

- In a standard system, precision decreases as either the number of docs retrieved or recall increases
  - This is not a theorem, but a result with strong empirical confirmation

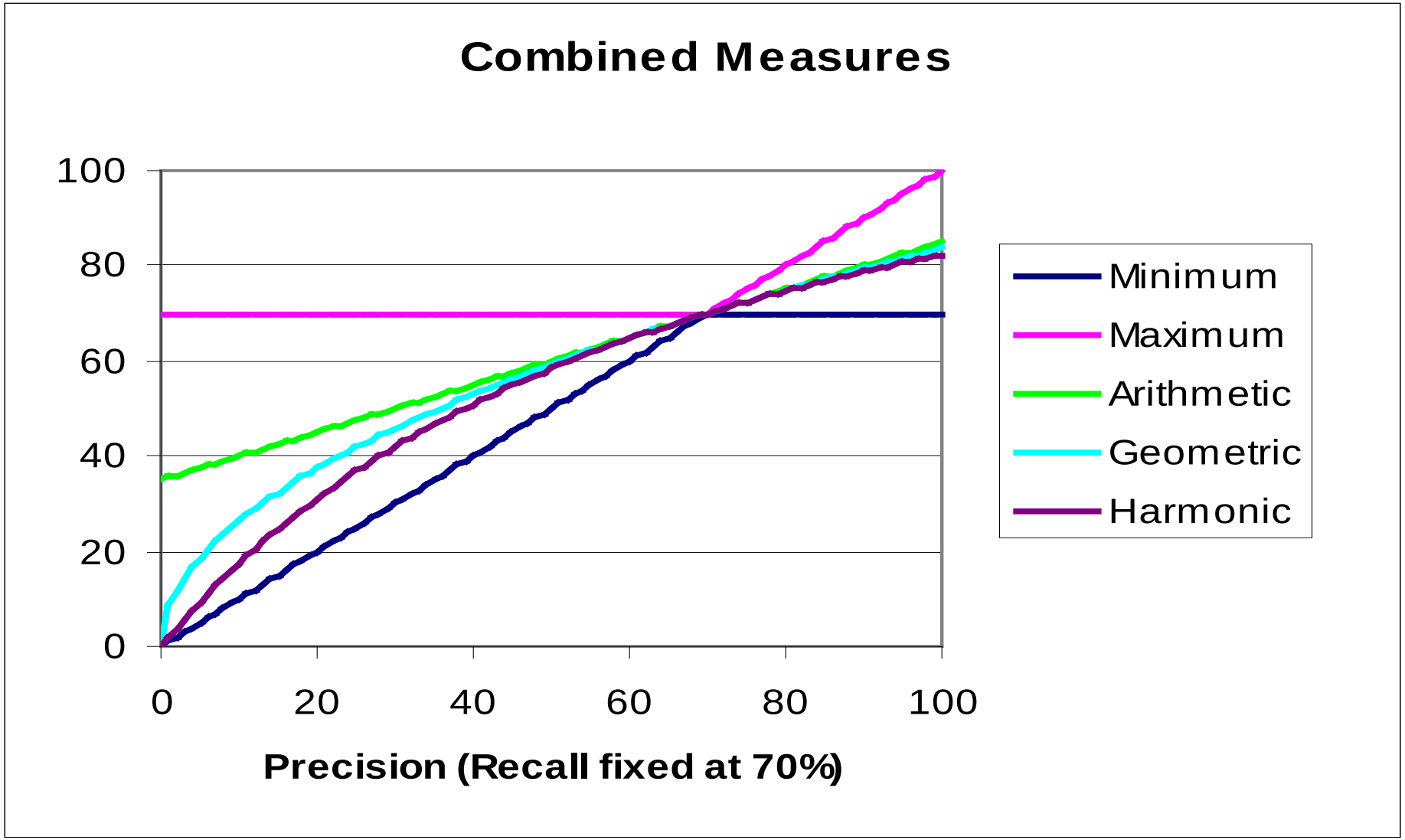
F-Score

- Combined measure that assesses precision/recall tradeoff is **F-score**:

$$F - Score = \frac{2 * precision * recall}{precision + recall}$$

- Harmonic mean is a conservative average
- *What does F stand for?*

Harmonic and Other Means



Evaluation of Ranked Lists

5 Ways to **Make** a Flying Model **Plane** from Scratch - wikiHow
www.wikihow.com › ... › Model Making › Model Aircraft
How to Make a Flying Model **Plane** from Scratch. Sure, **making** a model of an F-22 is fun, but as James Dicky once said "Flight is the only truly new sensation ...

Mini **airplane** - Instructables
www.instructables.com/id/mini-**airplane**/
Very **easy to make**, electric motor, wires and battery.... a flying **airplane**. ... The kind you used to be able to buy at **toy** stores and most likely now only at airport gift ...

How to Build a Battery Powered **Plane** (Balsa Wood **Airplane**)
www.instructables.com/.../**How-to-Build**-a-Battery-Powered-**Plane**-Balsa-...
A how to guide showing **how to make** your own battery powered balsa wood ...
Tags:how to **Toys** planes glider **airplane** aircraft flying hobby **toy plane** model ...

HOME MADE **Toy Airplane** (Glider) - YouTube


www.youtube.com/watch?v=7d-c0tLJDbI
Aug 30, 2013 - Uploaded by MINECRAFTIORD808
I have made a Wooden **Toy Airplane** or Glider! Sorry... I have no Tutorial BUT, You ... Website to **Make Plane** ...

AIRPLANE INTRO - Science **Toy** Maker
science**toymaker**.org/**plane**/index.htm
build a rubber band powered balsa **plane**. ... All my 6th grade students-- over 250 of them every year-- **make** this project from scratch. Thin strips of balsa wood ...

How to Build a Toy Airplane That Flies | eHow
www.ehow.com › Hobbies, Games & Toys
How to Build a Toy Airplane That Flies. To built a toy airplane that flies, the best place to start is with materials that are already around the house. This airplane is ...

How to Build a Toy Motorized **Airplane** | eHow
www.ehow.com › Hobbies, Games & Toys
Jun 2, 2014 - **How to Build a Toy** Motorized **Airplane**. Most kids know **how to make** paper airplanes by the time they reach the first grade. This article, however ...

Free Flight **Plane** using "DC **Toy** Motor" - RC Groups
www.rcgroups.com/forums/showthread.php?t=1208229
Mar 11, 2010 - 15 posts - 5 authors
Discussion Free Flight **Plane** using "DC **Toy** Motor" Free Flight. ... Hi I want to **build** a

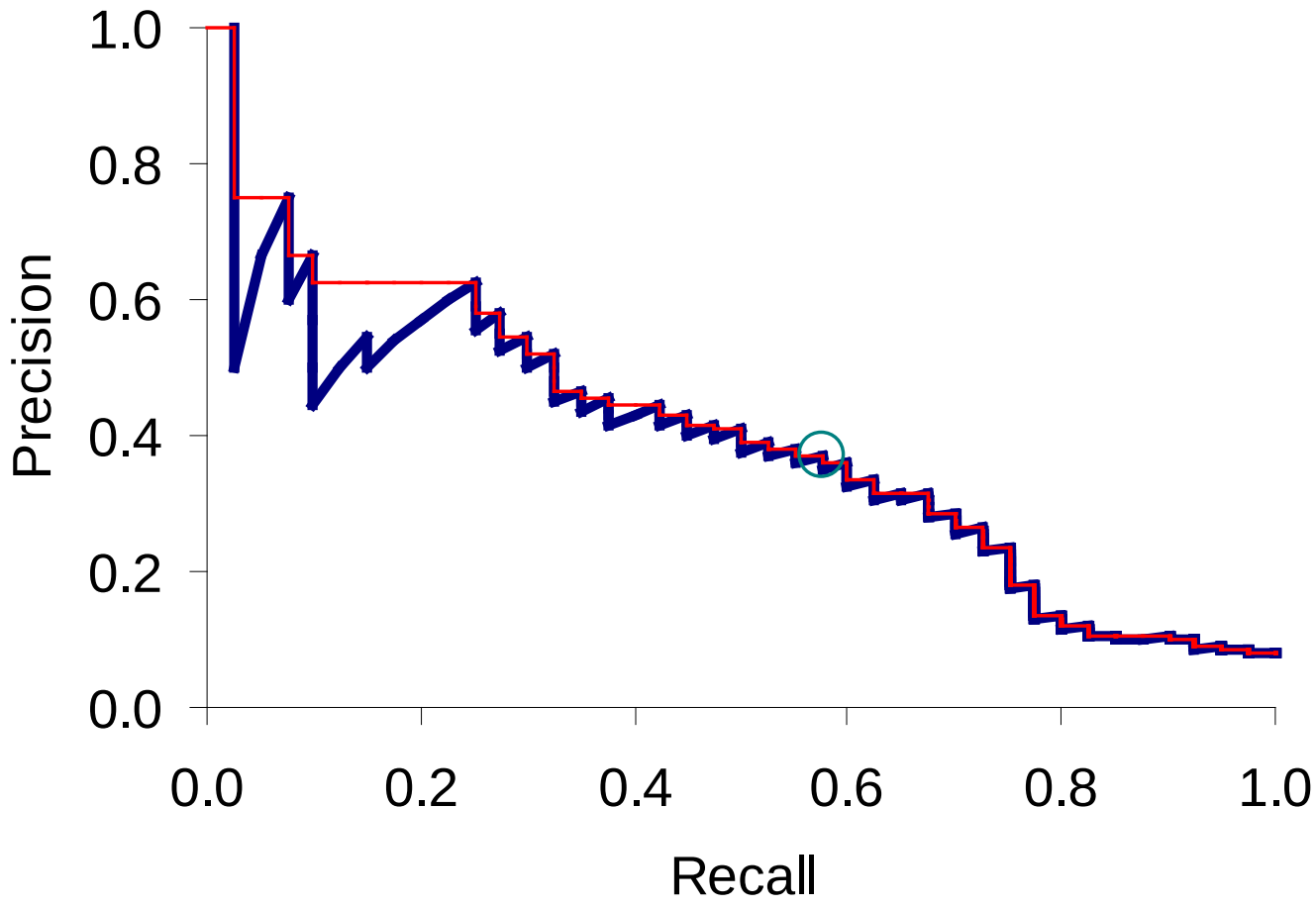
Relevant	Relevant	How To Make A Toy Plane - Tumblr howtomakeatoyplane.tumblr.com Before go into more details on **how to make a toy plane**, it is important to have a look at on how the history of airplanes changed. Having understanding of the ...
Non-Relevant	Relevant	How To Make A Toy Plane - blogspot.com howtomakeatoyplane.blogspot.com **How to make a toy plane** with easy steps | **How To Make A Toy Plane** Try your self and **make a toy plane** your own. It is easy and fun! Pages. Home; Video Articles ...
Non-Relevant	Relevant	YouTube www.youtube.com/watch?v=v207tVlVeKg If the owner of this video has granted you access, please log in.
Relevant	Non-Relevant	How To Make A Toy Plane: Rubber Band Powered Toy Plane howtomakeatoyplane.blogspot.in/2010/01/rubber-band-powered-toy... **How to make a toy plane** with easy steps | **How To Make A Toy** ... Before move into implementation of **toy plane** let's try to understand the concept behind the flying.
Non-Relevant	Non-Relevant	How to fabricate a remote-controlled toy plane for beginners? soni2006.hubpages.com › Games, **Toys**, and Hobbies **How to make** homemade remote control car plus electric RC cars? ... hey i wanna **make** a flying **toy** machine....i dont knw **how to make** it ...
Non-Relevant	Relevant	How to make an RC plane in Mr. Herbert's Science class ... www.youtube.com/watch?v=ccD9zCBMeS0 By NightFlyer · 10 min · 6,93,904 views · Added 02-10-2007 **How to make** an RC **plane** in Mr. Herbert's Science class. ... leftover parts were used to **make** a neat RC **airplane** as demonstrated in Mr. Herbert's Science Class.
Relevant	Non-Relevant	How to Make a Toy Airplane Model | eHow UK www.ehow.co.uk › Hobbies **How to Make a Toy Airplane** Model. **A toy aeroplane** flies by imagination, not with batteries or rocket fuel. ... **How to make** model **airplane** propellers from plastic cups.
Relevant	Non-Relevant	Airplane Crafts for Kids: Ideas to make Planes & Paper ... www.artistshelpingchildren.org/**airplanes**-craftsideasactivitieskids... Another fun **toy airplane** to **make** is one that is more just for pretend...ofr young kids such as preschoolers and other young kids.

Adobe

Evaluation of Ranked Lists

- Evaluation of ranked results
- The system can return any number of results
- By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a precision-recall curve

Precision-Recall Curve

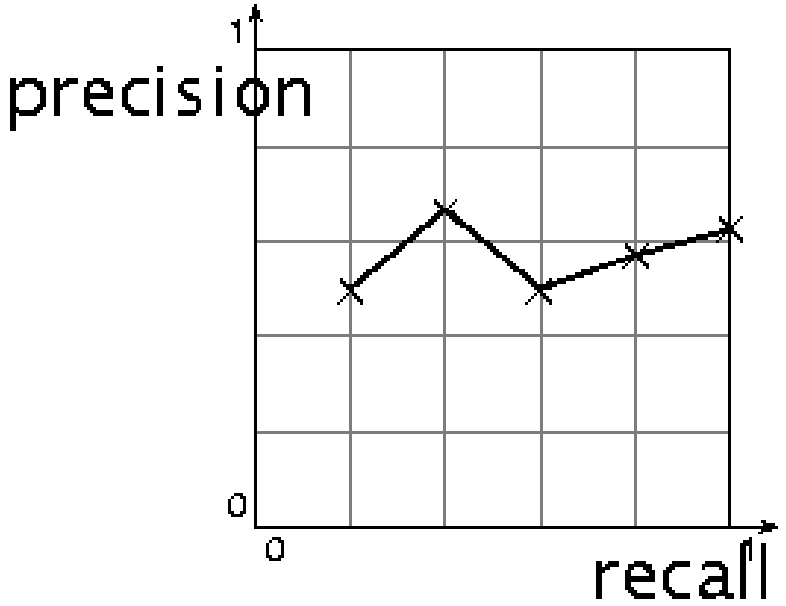


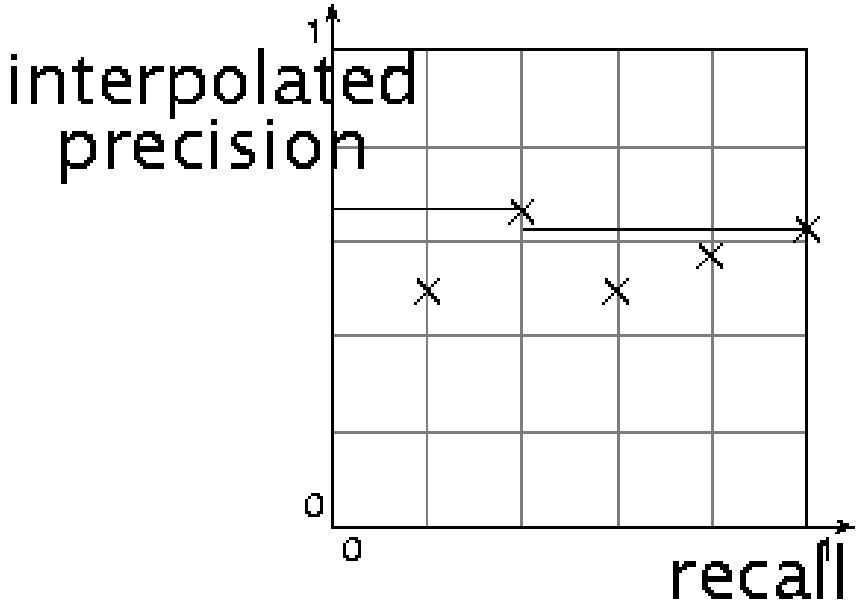
Averaging over queries

- A precision-recall graph for each query is not a very practical thing
- We need an aggregate performance indicator over a set of queries
- But there's a technical issue:
  - Precision-recall calculations place some points on the graph
  - How do you determine a value (interpolate) between the points?
  - Scatter plot to line graph

Interpolated Precision

- Idea: If locally precision increases with increasing recall, then we should give credit to the system for that
- So you take the max of precisions to right of value





P@k

- Graphs are good, but people want summary measures!
  - Precision at fixed retrieval level
    - Precision-at-*k* *(P@k)*: Precision of top *k* results
    - Perhaps appropriate for most of web search: all people want are good matches on the first one or two results pages
    - Averages badly and has an arbitrary parameter of *k*
    - Usually, $k = 1, 3, 5, 10, 20, 30, \ldots$

Average Precision

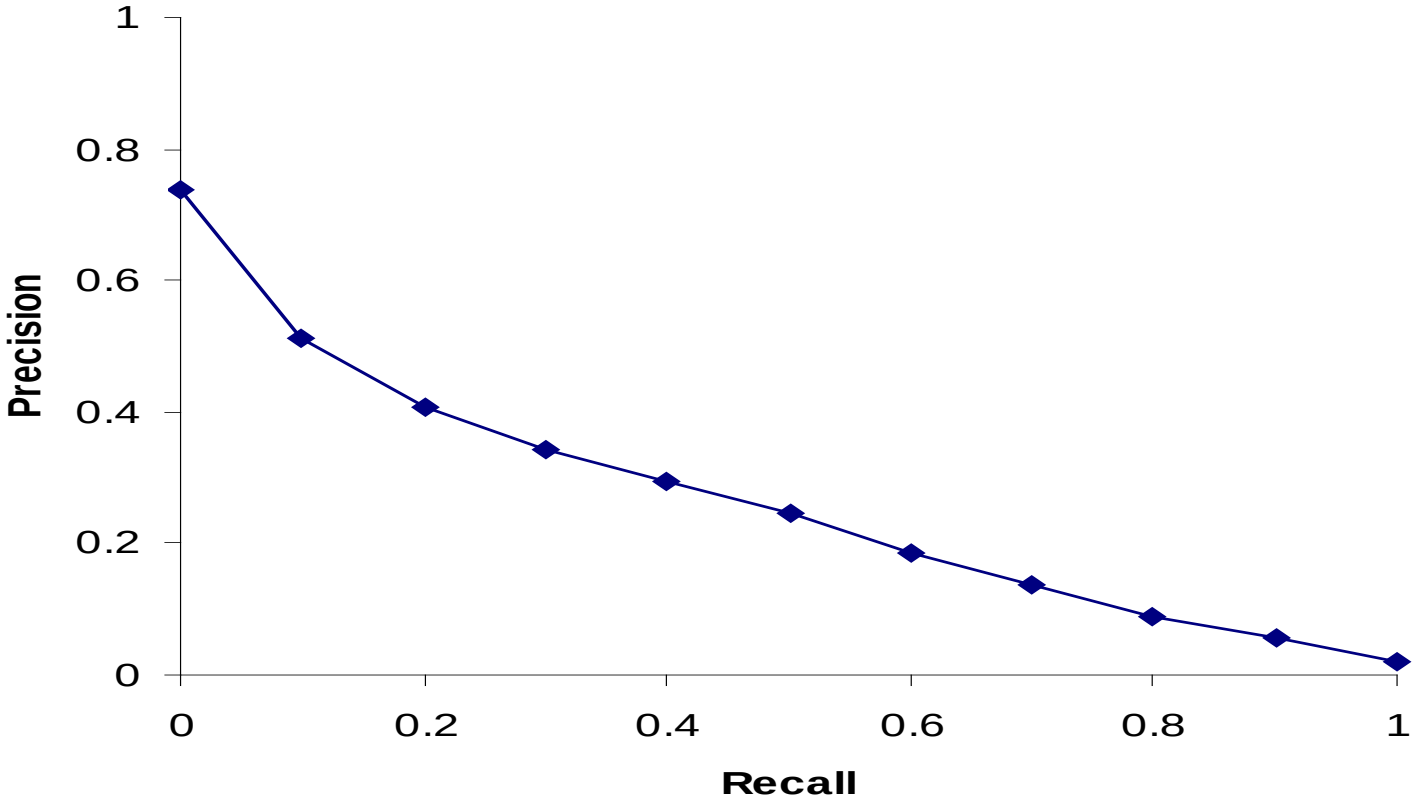
- More robust metric: Average Precision (AP)

$$Average\ Precision(q) = \frac{\sum_{k=1}^{n}(P@k \times rel(k))}{No.\ of\ relevant\ documents}$$

- *Rel(k)* is 0 or 1 according as the document is non-relevant or relevant
- **Mean Average** Precision (MAP) is the average AP over all queries q in the set Q

$$MAP(Q) = \frac{\sum_{q=1}^{Q} AP(q)}{|Q|}$$

Good (Typical) 11-point precision



Variance

- For a set of queries, it is usual that a system does badly on some information needs (e.g., MAP = 0.1) and excellently on others (e.g., MAP = 0.7)
- It is usually the case that the variance in performance of the same system across queries is much greater than the variance of different systems on the same query
- That is, there are easy information needs and hard ones!
- Automatic prediction of difficult queries

Mean Reciprocal Rank: Queries with "Correct" Answers

- Considers the rank of the first correct (relevant) answer

$$MRR(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{Rank_i}$$

- A good system would have the first relevant result at a higher rank
- Popular in question-answering systems

Graded Relevance Judgments: nDCG

- In several scenarios, judgments are not binary but graded
- Relevant, **partially relevant**, Non-relevant documents
- Excellent, Very good, good, fair, poor, detrimental
- Most popular metric with graded relevance judgments: Normalized Discounted Cumulative Gain (nDCG)

Graded Relevance Judgments: nDCG

- **Cumulative** Gain is the simple sum of relevance judgments, $CG_p = \sum_{i=1}^{p} rel_i$
- **Discounted** Cumulative Gain, $DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 p}$
- **Ideal** Discounted Cumulative Gain, $IDCG_p$ is the DCG_p for an ideal ranked list
- **Normalized** Discounted Cumulative Gain, $nDCG_p = \frac{DCG_p}{IDCG_p}$
- Average over query set also called nDCG, not MnDCG ☺

Metrics try to model user behavior

- All metrics try to capture some user model of relevance (happiness)
- Higher proportion of relevant results
- More coverage of relevant results
- Ranked list is easier to navigate
- Results at top attract more attention
- User attention drops non-linearly with rank

Evaluation at Large Search Engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the Web
- Search engines also use non-relevance-based measures
  - Clickthrough on first result
  - Clicked, skipped, missed URLs
  - First clicked, last clicked URLs
  - Text selections and mouse overs
- A/B testing

Evaluation at Large Search Engines

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness
- Probably the evaluation methodology that large search engines trust most

Why evaluation is important

- It allows you to convince others (e.g., reviewers, other researchers, funding agencies) that your work is meaningful
- Without a strong evaluation, your paper will (probably) be rejected
- Empirical evaluation helps guide meaningful research directions

Key steps

- What do we evaluate?
- Before writing any code, running any experiments, etc., it is essential to:
  - Clearly define a *task*
  - Formulate a set of *testable hypotheses*

Defining tasks

- It is very difficult to solve ill-defined problems/tasks
- Are you solving a well-known existing task? Or something that is similar to an existing task?
  - Ad hoc (query-based) retrieval, categorization, question answering, clustering
- Is your task novel and unlike anything that has been done before?
  - How is the system used?
  - What are the inputs? outputs?
  - How do you define success?
- Important: No matter how good your research is, or how good your results are, it is of little value if nobody understands what problem you're trying to solve

Formulating hypotheses

- Statement of the hypothesis to be tested
- The hypothesis either holds or does not, with some caveats:
- Scope: With respect to the data we use for experiments
- Extent: Might hold to a limited extent and/or under certain conditions
- Important questions to be asked:
- What specific component of our proposed method actually concerns the hypothesis? Can we test this component in isolation?
- Methods are not devised in an arbitrary manner. We always have a hypothesis (whether implicit or explicit) for why our work should improve upon existing research

Core Steps of Empirical Evaluation

- How do we evaluate?
- Key components necessary to carry out a meaningful evaluation:
- Experimental *methodology*
- Analysis of *results*

Experimental Methodology

- Provide full details of the implementation for reproducibility
- Free parameters, data processing (e.g., stopword removal, stemming), toolkits used, etc.
- Benchmarks
- Publicly available vs. proprietary
- Reference comparisons (baselines)
- Pertaining to the task (and hypothesis) at hand, strong enough

Experimental Methodology

- Re-implementation vs. quoting numbers
- Evaluation metrics
- Parameter tuning
- Tune parameters of new approach and reference systems
- Consider other factors that affect performance
- Key takeaway: If you're lucky, someone else may want to re-run your experiment in the future. It is up to you to make sure this is possible by using a good methodology and describing it adequately.

Analysis of Results

- How well did our methods do (on average) with respect to the reference comparisons?
- Are the improvements consistent across benchmarks?
- Anecdotal differences vs. substantial differences vs. *statistically significant* differences
- Analysis of factors affecting the performance, sensitivity to values of free parameters
- Analyzing the performance of specific cases (instances) of the success of our method, failure analysis is equally important
- *Bottom line: Do the results provide substantial support to the stated hypothesis?*

The Train/Development/Test Paradigm

- Consider the document retrieval task
- Your (hypothetical) hypothesis:
  - All query words must match in document
  - Longer documents are more useful
  - Documents with less words in title are more useful
  - Documents with more unique words in them are more useful

The Train/Development/Test Paradigm

- What are your features?
  - Document length (in number of words) – Feature 1 - *d*
  - Title length (in number of words) – Feature 2 - *t*
  - No. of unique words in document – Feature 3 – *u*
- How do you combine them?
  - How does the result vary with each feature?
  - Come up with a **model**

The Train/Development/Test Paradigm

- **Possible models:** d + u – t, (d * u)/t, (log d + log u)/t
- Is this enough?
- Perhaps different features have different weights: *α, β, γ*
- *αd + βu – γt, (αd * βu)/γt, (log αd + log βu) / γt*
- *You need a benchmark corpus and one number to evaluate your performance*
- *How do we proceed?*

The Train/Development/Test Paradigm

- Do we use entire dataset for **tuning** and evaluation?
  - Model tuning, Parameter tuning
- Using entire dataset may result in overfitting and loss of generalization
  - Fairness?
- Split data into three parts:
  - **Training, Development/Validation, Test Sets**
  - Rule of thumb: 60-20-20
  - Several random splits

The Train/Development/Test Paradigm

- *Use training data for tuning your model*
- *Use development data for tuning your parameters*
- *Use test data for tuning nothing – just report results!!*
- In absence of tunable parameters, use 70:30 split of training and test data (rule of thumb, depends on data availability, specific use case)
- Average results over multiple random splits wherever applicable
- Availability of data for shared task

The Train/Development/Test Paradigm

- The Train/Dev/Test Paradigm for Setting Free-Parameter Values
- Preferably, the train, dev and test queries should be for the same corpus
- If possible, avoid non-random splits of the query set
- The metric for which performance is optimized in the training and development phases should be the same as that used for evaluation over the test set
- Line/exhaustive search or grid search

Evaluating Experimental Results

- Now that you have some results, what do you do with them?
- Evaluation involves measuring, comparing, and analyzing your results
- Helps prove (or disprove) your hypotheses
- Demonstrates how your methods or systems compare against the existing research
- Provides fundamental insights into the underlying research problems being addressed

Evaluating Experimental Results

- Up until this point, we've discussed how to measure and compare performance
- It is tempting to stop here and declare victory (or defeat)
- By analyzing your empirical results deeper, you can often gain more insights into the problems being studied
- What are some deeper analysis techniques?
  - Parameter sensitivity
  - Statistical significance
  - Examine special cases

Evaluating Experimental Results

- You should now be able to:
- Devise strong experimental methodologies
- Convince yourself and others that your new system or method does (or does not) advance the state-of-the-art
- Evaluate other researchers' empirical evaluations

Summary of Idea

- Define the task
- State hypothesis
- Devise an experimental methodology

Setup

Experiment

- Run reference comparisons
- Run new method
- Tune parameters
- Test special cases

Evaluation

- Compute metrics
- Statistical testing
- Deeper analysis
- Evaluate hypothesis

Key Takeaways

- Empirical evaluation drives information retrieval innovation
- Experimental methodologies must be carefully devised with the scientific method
- Generalization is very important, so parameters should be carefully using held-out data
- Experiments should be reproducible, so experiments should use standard, publicly available data sets and open source IR toolkits whenever possible
- Reference comparisons are critical for convincing yourself and others of the utility of your research
- Detailed analyses often reveal deeper insights into results than average-case analyses

Sources

- Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze, *Introduction to Information Retrieval*, Chapter on Evaluation

http://www.stanford.edu/class/cs276/handouts/lecture8-evaluation.ppt

- Donald Metzler and Oren Kurland, *Experimental Methods for Information Retrieval*, SIGIR 2012 Tutorial

https://iew3.technion.ac.il/~kurland/sigir12-tutorial.pdf


Adobe