Information Retrieval Evaluation

Nicola Ferro
@frrncl

Intelligent Interactive Information Access (IIIA) Hub
Department of Information Engineering
University of Padua

ESSIR 2022, 13th European Summer School in Information Retrieval
18-22 July 2022, Lisbon, Portugal
Outline

- Basics of Experimental Evaluation
- Relevance Assessment
- Evaluation Measures
- Statistical Hypothesis Testing
- Foundations of Measurement
- Reproducibility
Evaluation Basics
Our Goal
Our Goal
Deeply Rooted…

Experimentation
Deeply Rooted…

Experimentation
Why Evaluation?

“To measure is to know”

“If you cannot measure it, you cannot improve it”

Lord William Thompson, first Baron Kelvin (1824-1907)
What to Evaluate?

Efficiency  vs  Effectiveness
Evaluation in Action

(c) IBM Corporation. http://www.youtube.com/watch?v=3G2H3DZ8rNc

IBM Watson: Deep QA Project
Evaluation in Action

IBM Watson: Deep QA Project

(c) IBM Corporation. http://www.youtube.com/watch?v=3G2H3DZ8rNc
Critical Issues in Evaluation

- It must be scientifically **valid**
  - valid methodology, measures, and statistics
  - large-scale enough to be statistically valid
  - must be “repeatable” if possible

- It must be **realistic** for the applications that will be using the information retrieval systems
  - **task** and use cases

- It must be **understandable** to your audience/client

Evaluation Spectrum

Fig. 2.1 Research continuum for conceptualizing IIR research.

This type of study differs from a pure system-centered study because researchers recruit users to make assessments and build new infrastructure, rather than relying on the TREC infrastructure. This is often done because researchers are working on new problems or tasks that have not been addressed by TREC. For example, Teevan et al. [269] studied relevance feedback and personalization; this required the collection of queries, documents, and relevance assessments from users. Although it is possible to study the interaction between the user and the information need, or the user and documents, this is usually not the focus of this type of study.

Intent and purpose of the research are important in determining where a study belongs along the continuum. Consider a study where a system evaluation with users has been conducted but the researchers are primarily interested in demonstrating the goodness of the system, rather than understanding the user-system IR interaction; the user study is, in effect, an ancillary task rather than a central focus. In many ways, these types of studies undermine efforts to create a more solid foundation for IIR studies, since users are essentially treated as black boxes. Although it is not claimed that all IR studies should focus on users, an explicit mention of the focus of the study should be made so that readers can better distinguish between findings about IR systems, findings about interactive IR and findings about users. There is also

How Does Experimental Evaluation Work

- **Cranfield Paradigm** by Cyril W. Cleverdon
  - Dates back to mid 1960s

- Makes use of **experimental collections**
  - **documents** (corpora)
  - **topics**, which are a surrogate for information needs
  - **relevance judgments** (binary or graded)
    - also called relevance assessment or ground-truth (or qrels)

- Ensures **comparability** and **repeatability** of the experiments


Some Document Collections

- **Historical**
  - **CACM**: 3,024 abstracts from the Communications of the ACM  
    http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/

- **Mid-nineties**
  - **TIPSTER**: 528,155 documents (news articles, US government reports, ...), Disks 4 and 5 excluding Congressional Record subcollection  
    https://catalog.ldc.upenn.edu/LDC93T3A

- **Early 2000s**
  - **WT10g**: 1,692,096 Web pages crawled in 2001  
    http://ir.dcs.gla.ac.uk/test_collections/wt10g.html
  - **GOV2**: 25,205,179 Web pages crawled from .gov sites in early 2004  
    http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
  - **CLEF Multilingual Corpus**: 4,883,227 multilingual news articles corpus in 13 languages (Bulgarian, Dutch, English, Farsi, Finnish, French, German, Hungarian, Italian, Portuguese, Spanish) gathered in 1994, 1995 and 2002. Topics in 28 different languages (Bengali, Bulgarian, Chinese, Czech, Dutch, English, Farsi, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Marathi, Norwegian, Oromo, Polish, Portuguese, Russian, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai)

- **Today**
  - **ClueWeb 2009**: 1,040,809,705 Web pages in 10 languages crawled between January and February 2009  
    https://lemurproject.org/clueweb09/
    https://lemurproject.org/clueweb12/
    https://catalog.ldc.upenn.edu/LDC2008T19
  - **TREC Washington Post Corpus**: 595,037 news articles and blog posts from January 2012 through August 2017 from Washington Post  
    https://trec.nist.gov/data/wapost/
  - **MS MARCO**: 3.2 million English documents, 8.8 million passages, 1 million questions  
    https://microsoft.github.io/msmarco/
Evaluation with Test Collections in a Nutshell

<table>
<thead>
<tr>
<th>Topic</th>
<th>Run</th>
<th>Assessed Run</th>
<th>Weighted Assessed Run</th>
<th>Measure Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d_1$</td>
<td>Highly Relevant</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_2$</td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_3$</td>
<td>Partially Relevant</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_4$</td>
<td>Fairly Relevant</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_5$</td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_6$</td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

$DCG = 4.6309$

Evaluation with Test Collections in a Nutshell

Example of Topic

Topics consists of:

- **title**: a brief statement expressing the information need. It resembles the typical search engine query
- **description**: more detailed formulation of the information need
- **narrative**: instructions for assessors on when to consider a document relevant

Typical experimental collections make use of 50 topics
Evaluation with Test Collections in a Nutshell

<table>
<thead>
<tr>
<th>Topic</th>
<th>Run</th>
<th>Assessed Run</th>
<th>Weighted Assessed Run</th>
<th>Measure Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>d₁</td>
<td></td>
<td>Highly Relevant</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>d₂</td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>d₃</td>
<td></td>
<td>Partially Relevant</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>d₄</td>
<td></td>
<td>Fairly Relevant</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>d₅</td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>d₆</td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

$$DCG = 4.6309$$

Evaluation with Test Collections in a Nutshell


\[ DCG = 4.6309 \]
The media have a "tremendous problem," Meyerhoff says, in trying to accurately report on the Alar scare. "Reaction to the Alar scare "set a troubling precedent," the Washington Post editorialized several weeks later. "A complicated scientific issue was allowed to be decided not by officials charged with protecting the public, on the basis of hard evidence, but by a frightened public acting on incomplete and often erroneous press reports." The media coverage produced a nationwide hysteria. Indeed it did — three weeks later, right after “60 Minutes” aired its story. News media throughout the country — effectively manipulated by the Natural Resources Defense Council and aided by public appeals and congressional testimony from that well-known molecular biologist Meryl Streep — almost made it seem that one bite of an Alar-treated apple or one swig of juice made from Alar-treated apples would mean instant death.

Coverage of the Alar scare was "outrageous...completely alarmist," says Marla Cone, who writes about the environment for the Los Angeles Times.

But Alar was a made-to-order media story. It had apples, kids and cancer. "A lot of (media) people were suckerized," Cone says.

The media coverage produced a nationwide hysteria. School boards in Los Angeles, New York, Chicago, Atlanta and many other cities banned apples and apple products from their cafeterias. Some parents raced after their children’s school buses to yank apples from their lunch boxes. Supermarkets came under intense pressure to remove apples form their shelves. Uniroyal, the manufacturer of Alar, pulled the product off the market. Sales of apples plummeted, forcing many farmers to dump their crops or give them away — costing the industry more than $100 million, according to economists’ estimates.

Reaction to the Alar scare “set a troubling precedent,” the Washington Post editorialized several weeks later. "A complicated scientific issue was allowed to be decided not by officials charged with protecting the public, on the basis of hard evidence, but by a frightened public acting on incomplete and often erroneous press reports.”

The EPA had expressed concern about the safety of Alar for many years before Newsweek and “60 Minutes” jumped on the story. But the agency had decided that test results were either flawed, contradictory or insufficiently conclusive to warrant an immediate ban and formal action was delayed, pending hearings in late 1988.

David Gelber, the producer of the “60 Minutes” Alar story and now the executive producer for ABC’s “Peter Jennings Reporting,” says EPA and scientific criticism of Alar convinced him the story was worth doing at the time.

Dr. John A. Moore, then acting administrator of the EPA, said on “60 Minutes” that Alar “should come off the market” because of what he had earlier described as “an inescapable and direct correlation” between exposure to Alar and “the development of life-threatening tumors.”

“The public had a right to know that was their view,” Gelber says.

Al Meyerhoff, senior attorney for the Natural Resources Defense Fund, also defends the Alar story.

He says the apple industry launched a “concerted disinformation campaign” in an effort to persuade the media and consumers alike that Alar was not
The environmental community, which recently split deeply over support for the North American Free Trade Agreement, is issuing warnings about the new world trade agreement in a newly unified voice.

The environmentalists charge that the comprehensive General Agreement on Tariffs and Trade accord signed Friday in Morocco will erode the United States' -- and California's -- ability to enforce its environmental strictures on everything from recycling to pesticide use to air pollution.

"If this . . . is enacted," said Barbara Dudley, executive director of Grassroots USA, "over two decades of environmental protection could be severely weakened."

Nowhere is that more true than in the area of food safety, some environmentalists argue.

The United States, with some of the world's most restrictive regulation of pesticides, prohibits the entry of food products with detectable traces of about 48 chemicals -- substances used by many of its trading partners and listed as allowable by the standard-setting organization of the new agreement.

But if U.S. Customs Service inspectors begin turning away food imports that bear traces of these chemicals under the new agreement, America's trading partners are almost certain to cry foul, environmentalists warned. If their challenges stand, one activist said, U.S. pesticide protections could topple, one after another.

Supporters of the accord acknowledged that the letter of the new agreement may indeed put the United States, with its strong environmental protections, on the defensive. That is because some U.S. laws championed by environmentalists, as well as state laws, do not appear to be based on undisputed scientific evidence demonstrating that a regulation will improve the public's health or mitigate a known environmental hazard.

Supporters of the agreement also argue that the United States' market power, as well as a growing appreciation for the environmental ethic among America's trading partners, will cause virtually any challenge to U.S. environmental laws to fail.

"The practical implementation of trade agreements is often more politically sensitive and realistic than the sheer language of the treaty," said William K. Reilly, administrator of the U.S. Environmental Protection Agency in the George Bush Administration and an avowed free trader.

Other free traders cited the power and transparency of the processes by which U.S. laws are passed and regulations are made. In a challenge before the international trade deliberative body, U.S. defenders could produce reams of scientific data, risk assessments and economic analysis to prove that an American environmental standard was established not to keep foreign products out, but to benefit the public's health.

"In a fair fight," said Linda Fisher, a Washington-based trade attorney with the Los Angeles law firm Latham & Watkins, "the United States will win."

But environmental activists are not comforted by those assurances. They say that the fine print of the new agreement would allow a trading partner to argue before the trade court in Geneva, Switzerland, that Washington's environmental laws -- or those of individual U.S. states -- constitute an unfair barrier to the entry of that country's exports.

If the United States regulates more strictly than its trading partners the fuel emissions or stipulates the fuel efficiency of cars sold or operated within its borders (which it does), or prohibits food products that bear traces of certain pesticides it considers hazardous (which it does), a trading partner, in principle, can challenge the federal stricture under the trade agreement.

Environmentalists said that under the trade agreement, trading partners could target laws such as the one that grew out of California's Proposition 65, which requires a cautionary label on any product that would expose its user to a carcinogen or a chemical that could be harmful to a developing fetus or pregnant woman.

Other California state regulations that could be challenged go beyond federal government requirements by making manufacturers of agricultural chemicals furnish the state with data on the chemical's possible effects on human reproduction, water pollution, exposed workers and endangered species.

Another California stricture that could come under attack is one that has required wine manufacturers to sponsor efforts to warn consumers of the possible dangers posed by the lead in the foil that covers wine corks. Finally ... a World Trade Pact?

Ministers from 124 nations ended seven years of complex negotiations Friday and formally concluded the Uruguay Round of the General Agreement on Tariffs and Trade talks in Marrakesh, Morocco. Following are the main elements of the 10,000-page, 385-pound global world trade pact:

MARKET ACCESS -- This is the backbone of the act. Countries pledge to cut tariffs on industrial and farm goods by an average of about 37%. The United States and European Union agree to trim tariffs between them by one half.

SERVICES -- For the first time, rules will govern annual trade in services such as banking, insurance and travel, as well as the movement of labor. The United States reserves the right to deny other countries favorable access to its lucrative U.S. financial services market, but will hold off for at least 18 months. Washington has threatened to challenge EC curbs on audio-visual goods.
But at Little Saigon Supermarket, Ngo chooses from an increasing range of Vietnamese-style processed meats, all sorts of fresh noodles, herbs, pickled vegetables and soy products (such as fresh tofu and soy milk) that are impossible to import. The selection has made Vietnamese eating here as close to authentic as it gets outside of Vietnam itself.

As the selection shows, Little Saigon Supermarket owner David Tran knows the Vietnamese food business inside and out. Tran came to the Little Saigon area at the age of two. When it was still in its awkward growth stages — a mere stretch of bean and strawberry fields, flower warehouses and bottling plants, and just a few Vietnamese stores and businesses — Tran and his family settled here in Orange County near the neighborhood centered on Bolsa Avenue — and running from Westminster through Garden Grove to Santa Ana — that came to be known as Little Saigon. Lured by the warmer climate and a growing, cohesive Vietnamese community, the Nguys found many ties to their culture here, one of the strongest being the ready availability of Vietnamese foods and ingredients.

"You won’t find most of these vegetables in Wisconsin," Ngo says, picking over the greens in the produce section of the bright, ultra-modern Little Saigon Supermarket on Bolsa Ave. By “these vegetables,” she means the dozens of Vietnamese specialties such as bò qua, the pale green spongy stems that go into your soup, or mếp hùng, a squash resembling a large zucchini.

In recent years, specialty produce farming has turned into a livelihood for a number of local Vietnamese farmers, evolving hand in hand with a new-sizeable Vietnamese food industry. Many of these California-grown or -made goods are distributed nationwide to serve the million or so Vietnamese who have come to the United States over the years.

At the center of this commercial activity is Orange County, where Vietnamese residents officially increased 271% in the last decade (the Vietnamese-American Political Action Committee contends the number is nearly twice what the census reports). The sheer size of this local customer base opened up a lucrative market for prepared foods as well as shelf-stable items; in much smaller Vietnamese communities, it probably wouldn’t be economic to market these perishables.

The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

New, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.

"At the time, there’s as much competition among the various brands of these foods as there is between American breakfast cereal or coffee companies. You see five or six styles of curry powder and at least half-a dozen make the bologna-like Vietnamese sausages chua lua and cha bo. Everyone has a long list of seasoning mixes in its line. The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

Now, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.

The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

Now, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.

The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

Now, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.

The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

Now, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.

The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

Now, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.

The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

Now, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.

The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on food they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren’t created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.

Now, however, Vietnamese in this country are no longer dependent on imports. They’re producing their own extravagant assortments of Vietnamese-style soups, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.
Where Do Topics and Collections Come From?

**Topics**
- Collection exploration
- Logs
- Observation of real users

**Collections**
- Opportunistic, e.g. my email or university Web site
- Constructed, e.g. tweets with #hashtag or results of a query to a search engine
- Naturalistic, e.g. large Web crawl, a year of news, a month of tweets

---

Evaluation with Test Collections in a Nutshell

<table>
<thead>
<tr>
<th>Topic</th>
<th>Run</th>
<th>Assessed Run</th>
<th>Weighted Assessed Run</th>
<th>Measure Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DCG = 4.6309</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Highly Relevant</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Partially Relevant</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fairly Relevant</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Evaluation with Test Collections in a Nutshell

<table>
<thead>
<tr>
<th>Topic</th>
<th>Run</th>
<th>Assessed Run</th>
<th>Weighted Assessed Run</th>
<th>Measure Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$DCG = 4.6309$</td>
</tr>
<tr>
<td>$d_1$</td>
<td></td>
<td>Highly Relevant</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>$d_2$</td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$d_3$</td>
<td></td>
<td>Partially Relevant</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$d_4$</td>
<td></td>
<td>Fairly Relevant</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>$d_5$</td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$d_6$</td>
<td></td>
<td>Not Relevant</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Example of Ground-truth (trec_eval format)

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Fixed</th>
<th>Document ID</th>
<th>Judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>0</td>
<td>LA050394-0237</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>LA112394-0177</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>LA091294-0164</td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>LA040594-0187</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>LA041694-0248</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0</td>
<td>LA031694-0234</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>0</td>
<td>LA040494-0111</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>0</td>
<td>LA081794-0171</td>
<td>1</td>
</tr>
</tbody>
</table>

- Relevance judgements (qrels) are textual files whose field are separated by tab or space
- Typically, for each topic there are 300-700 judgement documents and the number of judged document vary from topic to topic
Evaluation with Test Collections in a Nutshell

Evaluation with Test Collections in a Nutshell

<table>
<thead>
<tr>
<th>Topic</th>
<th>Run</th>
<th>Assessed Run</th>
<th>Weighted Assessed Run</th>
<th>Measure Score</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>d1</th>
<th>![Image of documents]</th>
<th>Highly Relevant</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>d2</td>
<td>![Image of documents]</td>
<td>Not Relevant</td>
<td>0</td>
</tr>
<tr>
<td>d3</td>
<td>![Image of documents]</td>
<td>Partially Relevant</td>
<td>1</td>
</tr>
<tr>
<td>d4</td>
<td>![Image of documents]</td>
<td>Fairly Relevant</td>
<td>2</td>
</tr>
<tr>
<td>d5</td>
<td>![Image of documents]</td>
<td>Not Relevant</td>
<td>0</td>
</tr>
<tr>
<td>d6</td>
<td>![Image of documents]</td>
<td>Not Relevant</td>
<td>0</td>
</tr>
</tbody>
</table>

## Example of Run (trec_eval format)

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Fixed</th>
<th>Document ID</th>
<th>Rank</th>
<th>Score</th>
<th>Run ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>Q0</td>
<td>LA050394-0237</td>
<td>1</td>
<td>0.6342</td>
<td>updrun</td>
</tr>
<tr>
<td>41</td>
<td>Q0</td>
<td>LA091294-0164</td>
<td>2</td>
<td>0.4278</td>
<td>updrun</td>
</tr>
<tr>
<td>41</td>
<td>Q0</td>
<td>LA040594-0187</td>
<td>3</td>
<td>0.4278</td>
<td>updrun</td>
</tr>
<tr>
<td>41</td>
<td>Q0</td>
<td>LA041694-0248</td>
<td>4</td>
<td>0.3197</td>
<td>updrun</td>
</tr>
<tr>
<td>41</td>
<td>Q0</td>
<td>LA102394-0113</td>
<td>5</td>
<td>0.3005</td>
<td>updrun</td>
</tr>
</tbody>
</table>

...  

| 42       | Q0    | LA081794-0171 | 1    | 0.7687 | updrun  |
| 42       | Q0    | LA031694-0235 | 2    | 0.7011 | updrun  |
| 42       | Q0    | LA031694-0234 | 3    | 0.6950 | updrun  |

- Runs are textual files whose field are separated by tab or space.
- Typically, there are 50 topics and 1,000 documents are retrieved for each topic (i.e. 50,000 lines)
Large-scale Evaluation Initiatives: TREC

- **TREC** (Text REtrieval Conference), USA, since 1992
- [https://trec.nist.gov/](https://trec.nist.gov/)

---

Large-scale Evaluation Initiatives: NTCIR

NTCIR (NII Testbeds and Community for Information access Research), Japan, since 1999


Noriko Kando

Large-scale Evaluation Initiatives: CLEF

**CLEF** (Conference and Labs of the Evaluation Forum), Europe, since 2000

http://www.clef-initiative.eu/

---

The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) is a self-organized body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure. CLEF promotes research and development by providing an infrastructure for:

- multilingual and multimodal system testing, tuning and evaluation;
- investigation of the use of unstructured, semi-structured, highly-structured, and semantically enriched data in information access;
- creation of reusable test collections for benchmarking;
- exploration of new evaluation methodologies and innovative ways of using experimental data;
- discussion of results, comparison of approaches, exchange of ideas, and transfer of knowledge.

The CLEF Initiative is structured in two main parts:

1. a series of Evaluation Labs, i.e. laboratories to conduct evaluation of information access systems and workshops to discuss and pilot innovative evaluation activities;
2. a peer-reviewed Conference on a broad range of issues, including:
   - investigation continuing the activities of the Evaluation Labs;
   - experiments using multilingual and multimodal data; in particular, but not only, data resulting from CLEF activities;
   - research in evaluation methodologies and challenges.

Since 2000 the CLEF has played a leading role in stimulating investigation and research in a wide range of key areas in the information retrieval domain, becoming well-known in the international IR community. It has

---

Large-scale Evaluation Initiatives: FIRE

**FIRE** (Forum for Information Retrieval Evaluation), India, since 2008

http://fire.irsi.res.in/
Evaluation Initiatives: Typical Cycle

The “Ideal Test Collection” Today

**:Corpora → (not historical) corpora are typically OK**
- < 500 documents, no real value
- 1-2,000 documents, minimally acceptable
- > 10,000 documents, actually needed

**:Topics → typical size is still 50 topics**
- < 75 topics, no real value
- 250 topics, minimally acceptable
- > 1,000 topics, actually needed

**:Relevance Judgements → binary is still most common option, diversity only recently**
- multi-graded (highly and fairly relevant)
- types (novel, stimulating, …)
- need for pooling (still open research issue)

Karen Spärck Jones

C. J. “Keith” van Rijsbergen

---

How Valuable is Evaluation?

- The **TREC 2010 Economic Impact** study estimated in about **30 M$** the overall investment in TREC by NIST.
  - probably much much more if we had a means to estimate also the investment by participants in TREC

- They are the **pillars** for all the subsequent scientific research and technology development.

- TREC estimated the **return on investment** in the range of **3$-5$** for each invested dollar.

---

Questions?

There's a problem with the nut...

EVALUATION TOOL
Relevance Assessment
Traditional Depth-k Pools

Relevance Assessment

2 judgments per minute
75 person/days per pool
35,000-75,000 documents per pool


What Makes a Good Pool?

- Depth-\(k\) pools require large enough \(k\) and different enough pooled systems in order to produce “complete” judgements.
  - Not pooled/not assessed documents are typically assumed to be not relevant.
  - This is a motivation for organizing large-scale evaluation initiatives.
- The objective is not to allow for computing the “exact” value of an evaluation measure but rather to **comparatively assess systems** and detect significant differences in a robust way.

**Leave-one-out tests:** are used to assess the **reusability** of a pool.
  - one system/group of systems is removed from the pool.
  - all the systems are evaluated using both the original pool and the newly created one.
  - the two sets of results are compared by computing the Kendall’s \(\tau\) correlation among the ranking of systems on the original and the new pool and/or the maximum drop in ranking.
Shallow Pools

Internal test collections used by commercial search engines have large numbers of topics (ten of thousands), much more than existed in publicly available ones.

The TREC Million Query Track created a test collection with 1,755 topics (using stratified sampling techniques).

250 topics with 20 judgments per topic are the most cost-effective in terms of minimizing assessor effort and maximizing accuracy in ranking runs.

There might be concerns about how much reusable are these test collection.

Need to develop ad-hoc evaluation measures aware of the sampling procedures.

Multi-armed Bandits Pools

- Bandit techniques trade-off between exploiting known good “arms” and exploring to find better arms.
  For collection building, each run is an arm, and reward is finding a relevant doc.

- Simulations suggest can get similar-quality collections as pooling but with many fewer judgments.

- TREC 2017 Common Core track first attempt to build new collection using bandit technique.


Slide courtesy of Ellen M. Voorhees (see her seminar)
Crowdsourcing

- Large scale, inexpensive, diverse
- Careful design of the task, attention to details, simplicity and usability
- Need assessment of quality of work

Crowdsourcing

Crowd Assessors

Aggregating Relevance Assessments: Upstream Approach

**Input**: a set of relevance assessments from each assessor

**Output**: a single set of relevance assessments, from which an evaluation measure is computed
How To Aggregate Relevance Assessments?

Majority Vote (MV)  
Expectation Maximization (EM)

What Can Go Wrong in Downstream Approaches?

- Out of 10 relevant documents in a pool, just 1 document has been wrongly labelled as not relevant
  - thus there is a **10% error** with respect to the whole pool
- Run₁ represents the case where the mis-labelled document is retrieved in ranks 1 to 5, while the other runs show what could have happened if it had been correctly labelled
  - P@5, i.e. precision at 5 retrieved documents, passes from 0% to 20%, so a **100% error**
  - AP, i.e. average precision, passes from 7.65% to 14.07%-22.96%, so a **45.61%-66.67% error**

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>P@5</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run₁</td>
<td>👻</td>
<td>👻</td>
<td>👻</td>
<td>👻</td>
<td>👻</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td></td>
<td>0.0000</td>
<td>0.0765</td>
</tr>
<tr>
<td>Run₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>0.2000</td>
<td>0.1407</td>
</tr>
<tr>
<td>Run₃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>0.2000</td>
<td>0.1463</td>
</tr>
<tr>
<td>Run₄</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>0.2000</td>
<td>0.1556</td>
</tr>
<tr>
<td>Run₅</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>0.2000</td>
<td>0.1741</td>
</tr>
<tr>
<td>Run₆</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>0.2000</td>
<td>0.2296</td>
</tr>
</tbody>
</table>
What Can Go Wrong in Downstream Approaches?

- Out of 10 relevant documents in a pool, just 1 document has been wrongly labelled as not relevant.
  - Thus there is a **10% error** with respect to the whole pool.

- Run_1 represents the case where the mislabelled document is retrieved in ranks 1 to 5, while the other runs show what could have happened if it had been correctly labelled.

- P@5, i.e. precision at 5 retrieved documents, passes from 0% to 20%, so a **100% error**.
- AP, i.e. average precision, passes from 7.65% to 14.07%-22.96%, so a **45.61%-66.67% error**.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Run_1</th>
<th>Run_2</th>
<th>Run_3</th>
<th>Run_4</th>
<th>Run_5</th>
<th>Run_6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>☠️</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>☠️</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>☠️</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>☠️</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>☠️</td>
<td>R</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>☠️</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P@5</td>
<td>0.0000</td>
<td>0.2000</td>
<td>0.2000</td>
<td>0.2000</td>
<td>0.2000</td>
<td>0.2000</td>
</tr>
<tr>
<td>AP</td>
<td>0.0765</td>
<td>0.1407</td>
<td>0.1463</td>
<td>0.1556</td>
<td>0.1741</td>
<td>0.2296</td>
</tr>
</tbody>
</table>

Mislabelled documents produce different errors for different runs and measures and these errors might be greater than the error on the pool itself.
Input: a set of evaluation measures computed according to the relevance judgements of each accessor

Output: an aggregated evaluation measure

The AWARE Framework

We compute the final score of a system as a weighted average of the scores computed according to the relevance judgements of each crowd-assessor.

The challenge is how to estimate the accuracy $a_k(t)$ to be assigned to each assessor.

Unsupervised estimator: the accuracy is proportional to the “distance” from prototypical random assessors.

$$\mathbb{E}\left[\mu(\hat{r}_t)\right] = \sum_{k=1}^{l} \mathbb{E}\left[\mu(\hat{r}_t) | W = W_k\right] a_k(t)$$
Randomness in Assessment?

- Relevance judgments is an intrinsically not deterministic process
  - subjectivity of the notion of relevance
  - inter-assessor agreement and crowd-sourcing
  - variation in assessor own notion of relevance

- Nevertheless, when we use relevance judgements in evaluation measures, we treat them as absolutely exact
Relevance as a Binomial Random Variable

- For each (topic, document) pair, let model relevance as a Binomial Random Variable $X \sim B(1, p)$ which assumes the value 1 with probability $p$ and the value 0 with probability $1 - p$
  - Since $E[X] = p$, $p$ roughly represents the amount of relevance of a document
- As a consequence, runs become sequences of random variables and evaluation measures become transformations of random variables

Applications
- removing the distinction between binary and multi-graded relevance
- robustness to incomplete information (remember that not assessed documents are assumed to be not relevant)
- robustness to inter-assessor agreement
- merging of crowd-assessors
- ...


... And of course we'll assess our progress along the way.

Will you be using an enhanced assessment methodology?

I hope that means something. All I did was string together some words I heard in the hallway.

Um... I'll be assessing... by measuring... and um...

I'd better get in on this.

I can't support this project until I see your advanced assessment methodology plan.

I'll have it in ten minutes, assuming you don't know what it's supposed to look like.

Very good.

I'll be in the shower trying to wash my soul.
Evaluation Measures
“Measure what is measurable and make measurable what is not”

Galileo Galilei (1564-1642)
# A Taxonomy of Evaluation Measures

<table>
<thead>
<tr>
<th>Binary Relevance</th>
<th>Multi-graded Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set-Based Retrieval</td>
<td>Not widely agreed generalizations of Precision and Recall</td>
</tr>
<tr>
<td>Precision (P)</td>
<td>Discounted Cumulated Gain (DCG)</td>
</tr>
<tr>
<td>Recall (R)</td>
<td>...</td>
</tr>
<tr>
<td>F-measure (F)</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank-Based Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision at Document Cut-off (P@k)</td>
</tr>
<tr>
<td>Recall at Document Cut-off (R@k)</td>
</tr>
<tr>
<td>R-Precision (Rprec)</td>
</tr>
<tr>
<td>Average Precision (AP)</td>
</tr>
<tr>
<td>Rank-Biased Precision (RBP)</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

Set-Based Retrieval and Rank-Based Retrieval are two approaches to information retrieval evaluation.

## Binary Relevance
- **Precision (P)**
- **Recall (R)**
- **F-measure (F)**

## Multi-graded Relevance
- Not widely agreed generalizations of Precision and Recall

## General Approaches
- Precision at Document Cut-off (P@k)
- Recall at Document Cut-off (R@k)
- R-Precision (Rprec)
- Average Precision (AP)
- Rank-Biased Precision (RBP)
- Discounted Cumulated Gain (DCG)
- ...
Set-based Measures: Precision, Recall and F-measure

- **Precision** is the proportion of retrieved documents that are actually relevant.
- **Recall** is the proportion of relevant documents actually retrieved.
- Together, Precision and Recall measure **retrieval effectiveness**, meant as the ability of a system to retrieve relevant documents while at the same time holding back non-relevant ones.
  - Maximizing Precision and Recall corresponds to optimal retrieval in the sense of the **Probability Ranking Principle**, i.e. ordering documents by their decreasing probability of being relevant, and creates a tight link between retrieval models and evaluation.
- **F-measure** is the harmonic mean of Precision and Recall, summarising them into a single score.

\[
P = \frac{|A \cap B|}{|B|} \quad R = \frac{|A \cap B|}{|A|}
\]

\[
F = 2 \frac{P \cdot R}{P + R} = \frac{2}{\frac{1}{P} + \frac{1}{R}}
\]

---

Set-based Measures: Example

$P = \frac{4}{10} = 0.40$

$R = \frac{4}{8} = 0.50$

$F = 2 \cdot \frac{\frac{4}{10} \cdot \frac{4}{8}}{\frac{4}{10} + \frac{4}{8}} = \frac{4}{9} = 0.44$

Assume $|A| = 8$ relevant documents in total

Lenient mapping to binary relevance degrees
Rank-based Measures: Precision and Recall

- **Precision at Document Cut-off:**
  
  \[ P(k) = \frac{1}{k} \sum_{n=1}^{k} r_n \]
  
  where \( r_k \in \{0, 1\} \) is the relevance degree of the n-th document

- **Recall at Document Cut-off:**
  
  \[ R(k) = \frac{1}{RB} \sum_{n=1}^{k} r_n \]
  
  where \( RB = |A| \) is the recall base, i.e. the total number of relevant documents

- **Rprec** is Precision computed at the recall base
  
  \[ R\text{prec} = P(RB) \]
**Rank-based Measures: Example of Precision and Recall**

Assume:
- $RB = 8$ relevant documents in total
- Lenient mapping to binary relevance degrees

### Run Assessed Run

<table>
<thead>
<tr>
<th>Topic</th>
<th>Run</th>
<th>Assessed Run</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d_1$</td>
<td>Highly Relevant</td>
</tr>
<tr>
<td></td>
<td>$d_2$</td>
<td>Not Relevant</td>
</tr>
<tr>
<td></td>
<td>$d_3$</td>
<td>Partially Relevant</td>
</tr>
<tr>
<td></td>
<td>$d_4$</td>
<td>Fairly Relevant</td>
</tr>
</tbody>
</table>

### Binary Weighted Assessed Run

<table>
<thead>
<tr>
<th>Topic</th>
<th>Run</th>
<th>Assessed Run</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d_1$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$d_2$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$d_3$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$d_4$</td>
<td>1</td>
</tr>
</tbody>
</table>

### Calculations

- $P(5) = \frac{3}{5} = 0.600$
- $R(5) = \frac{3}{8} = 0.375$
- $Rprec = P(8) = \frac{4}{8} = 0.500$
Assume $RB = 5$ relevant documents in total

The Precision-Recall curve has a typical saw-tooth shape

- We may have multiple Precision values for the same Recall value
- It is difficult to compare runs because they may not have the same Recall values
Interpolated Precision-Recall Curve

To interpolate Precision at standard Recall value $R_j$ we use the maximum Precision obtained for any actual Recall value $R$ greater than or equal to $R_j$

$$ iP@R_j = \max_{R \geq R_j} P@R $$
Standard Interpolated Precision-Recall curves exhibit a typical inverse relationship among Precision and Recall, indicating a trade-off between these two goals of effectiveness.

Rank-based Measures: Average Precision

\[ AP = \frac{1}{RB} \sum_{k \in R} P(k) = \frac{1}{RB} \sum_{n=1}^{N} \left( \frac{1}{n} \sum_{m=1}^{n} r_m \right) r_n = \]

\[ = \frac{rr}{RB} \cdot \frac{1}{rr} \sum_{k \in R} P(k) \]

where

- \( R \) is the set of the rank positions of the relevant retrieved documents
- \( rr = |R| \) is the total number of relevant retrieved documents
- \( N \) is the total number of retrieved documents, i.e. the length of the run
- The **Mean Average Precision (MAP)** is the mean of AP over a set of topics

Differently from the other measures, this mean has its own name since it is the most widely used single number to summarise the whole performance of a system

Rank-based Measures: Example of Average Precision

Assume
- \( RB = 8 \) relevant documents in total
- Lenient mapping to binary relevance degrees

\[ AP = \frac{1}{RB} \left( P(1) + P(3) + P(4) + P(8) \right) \]
\[ = \frac{1}{8} \left( 1 + \frac{2}{3} + \frac{3}{4} + \frac{4}{8} \right) = \frac{35}{96} = 0.36 \]
Area under the Precision-Recall Curve

The Area Under the Precision-Recall Curve (AUC) is an important indicator of the overall system effectiveness, summarising the trade-off between Precision and Recall.

\[ AUC = \int P@R \, dR = \sum_{k} P(k) \Delta R(k) \]
Computing the Area under the Precision-Recall Curve

\[
AUC = \sum_{n=1}^{N} P(n) \left( R(n) - R(n-1) \right) \quad \text{assuming } R(0) = 0
\]

Run 1

Interpolated Precision Recall Curve

\[
AUC = 0.00(0.00 - 0.00) + 0.50(0.20 - 0.00) + 0.66(0.40 - 0.20) + 0.50(0.40 - 0.40) + 0.60(0.60 - 0.40) + 0.50(0.60 - 0.60) + 0.42(0.60 - 0.60) + 0.50(0.80 - 0.60) + 0.55(1.00 - 0.80) + 0.50(1.00 - 1.00) = 0.5620
\]
\[ AUC = \sum_{n=1}^{N} P(n) \left( R(n) - R(n - 1) \right) \]

- When the n-th document is not relevant, \( R(n) \) is equal to \( R(n - 1) \) and their difference goes to zero.

- Therefore, we can sum only on \( R \), i.e. the set of the rank positions of the relevant retrieved documents.

\[ AUC = \sum_{k \in R} P(k) \left( R(k) - R(k - 1) \right) \]
Area under the Precision-Recall Curve and Average Precision

\[ AUC = \sum_{k \in R} P(k) \left( R(k) - R(k - 1) \right) \]

- Two adjacent rank positions differ just for one relevant document and thus

\[ R(k) - R(k - 1) = \frac{1}{RB} \sum_{n=1}^{k} r_n - \frac{1}{RB} \sum_{n=1}^{k-1} r_n = \frac{r_k}{RB} = \frac{1}{RB} \]

- Therefore AUC is equal to AP

  - this is one motivation of the importance of AP

\[ AUC = \frac{1}{RB} \sum_{k \in R} P(k) = AP \]
Rank-based Measures: Discounted Cumulated Gain

\[ DCG(k) = \begin{cases} 
\sum_{n=1}^{k} r_k & \text{if } k < b \\
DCG(k - 1) + \frac{r_k}{\log_b(k)} & \text{if } k \geq b
\end{cases} \]

\[ = \sum_{n=1}^{k} \frac{r_k}{\max(1, \log_b(k))} \]

where the base of the logarithm \( b \) indicates the patience of the user in scanning the result list.

- \( b = 2 \) is an impatient user
- \( b = 10 \) is a patient user

DCG naturally handles multi-graded relevance

DCG does not depend on the recall base

DCG is not bounded in \([0, 1]\)


Kalervo Järvelin

Jaana Kekäläinen
Rank-based Measures: Discounted Cumulated Gain

$$DCG(k) = \begin{cases} 
\sum_{n=1}^{k} r_k & \text{if } k < b \\
DCG(k - 1) + \frac{r_k}{\log_b(k)} & \text{if } k \geq b 
\end{cases}$$

where the base of the logarithm $b$ indicates the patience of the user in scanning the result list

- $b = 2$ is an impatient user
- $b = 10$ is a patient user

- DCG naturally handles multi-graded relevance
- DCG does not depend on the recall base
- DCG is not bounded in $[0, 1]$

Rank-based Measures: Discounted Cumulated Gain

\[
DCG(k) = \begin{cases} 
\sum_{n=1}^{k} r_k & \text{if } k < b \\
DCG(k - 1) + \frac{r_k}{\log_b(k)} & \text{if } k \geq b 
\end{cases}
\]

\[
= \sum_{n=1}^{k} \frac{r_k}{\max(1, \log_b(k))}
\]

where the base of the logarithm \( b \) indicates the patience of the user in scanning the result list.

- \( b = 2 \) is an impatient user
- \( b = 10 \) is a patient user

DCG naturally handles multi-graded relevance

DCG does not depend on the recall base

DCG is not bounded in \([0, 1]\)


Jaana Kekäläinen
Kalervo Järvelin
Rank-based Measures: User Models

- Rank-based evaluation measures, implicitly or explicitly, embed a **user model** comprising:
  - A **browsing model** that describes how a user interacts with results;
  - A **model of document utility**, describing how a user derives utility from individual relevant documents;
  - A **utility accumulation model** that describes how a user accumulates utility in the course of browsing.

- User models may be more or less **artificial** and may be more or less **correlated** with actual user behaviour and preferences.

- In the case of DCG:
  - A **browsing model**: user steps down the ranked results one-by-one, until s/he reaches the stopping rank k which is picked with a probability proportional to the log of the rank.
  - A **model of document utility**: user gains something from each relevant document, proportional to its relevance degree.
  - A **utility accumulation model**: user gains from all of the relevant documents from ranks 1 through k.

---

Rank-based Measures: Example of Discounted Cumulated Gain

Assume

- \( RB = 8 \) relevant documents in total
- An impatient user

\[
\text{DCG} = 3 + \frac{1}{\log_2(3)} + \frac{2}{\log_2(4)} + \frac{2}{\log_2(8)} = 5.2976
\]
Rank-based Measures: Normalized Discounted Cumulated Gain

- To normalize DCG in [0, 1], you need to compute the ideal run, i.e. the pool sorted in descending order of relevance, which represents the best retrieval possible and the maximum value of DCG

\[
nDCG(k) = \frac{DCG(k)}{iDCG(k)}
\]

- nDCG is given by the DCG of the run divided by the DCG of the ideal run
Rank-based Measures: Example of Normalized Discounted Cumulated Gain

Assume

- $RB = 8$ relevant documents in total
- An impatient user

\[
\begin{align*}
DCG &= 5.2976 \\
iDCG &= 10.1996 \\
nDCG &= 0.5194
\end{align*}
\]
The user starts from the top ranked document and with probability $p$, called persistence, goes to the next document or with probability $1 - p$ stops.

Typical values for $p$ are: 0.5 for impatient users, 0.8 for patient users, and 0.95 for extremely patient users.

$$RBP = (1 - p) \sum_{n=1}^{N} p^{n-1} r_n = (1 - p) \sum_{k \in \mathcal{R}} p^{k-1}$$

The user starts from the top ranked document and with probability \( p \), called persistence, goes to the next document or with probability \( 1 - p \) stops.

Typical value for \( p \) are: 0.5 for impatient users, 0.8 for patient users, and 0.95 for extremely patient users.

\[
RBP = (1 - p) \sum_{n=1}^{N} p^{n-1} r_n = (1 - p) \sum_{k \in \mathcal{R}} p^{k-1}
\]
Assume

\[ p = 0.8 \] a patient user

- Lenient mapping to binary relevance degrees

\[
RBP = (1 - 0.8) \left( 0.8^{1-1} + 0.8^{3-1} + 0.8^{4-1} + 0.8^{8-1} \right) = 0.4723
\]
Questions?

It works...It doesn’t....It works........

Just a hunch....maybe we do need a better way to measure results.....
Statistical Hypothesis Testing
The Problem: Are They Different?

\[ x = \begin{bmatrix}
  3.1929 \\
  0.6575 \\
  1.4155 \\
 -0.2043 \\
 -1.6054 \\
  1.1446 \\
  1.1033 \\
  0.7418 \\
  0.5035 \\
  1.5477
\end{bmatrix} \]

\[ y = \begin{bmatrix}
  0.7471 \\
  1.6357 \\
  1.0390 \\
  1.1421 \\
  1.5731 \\
  3.5389 \\
  1.9563 \\
  0.6754 \\
  2.6305 \\
  1.9505
\end{bmatrix} \]

\[ z = \begin{bmatrix}
  0.6408 \\
  2.6859 \\
  3.0818 \\
  3.1258 \\
  2.9028 \\
  2.4371 \\
  1.7648 \\
  3.2616 \\
  2.0997 \\
  1.0346
\end{bmatrix} \]

\[ \hat{\mu}_X = 0.8497 \]
\[ \hat{\sigma}^2_X = 1.5284 \]
\[ \hat{\mu}_Y = 1.6889 \]
\[ \hat{\sigma}^2_Y = 0.7890 \]
\[ \hat{\mu}_Z = 2.3035 \]
\[ \hat{\sigma}^2_Z = 0.8255 \]
What Did We Do?

\[ \mathbf{x} = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix} \]

\[ \mu_X = \mathbb{E}[X] \]
\[ \sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] \]
What Did We Do?

\[
x = \begin{bmatrix}
3.1929 & \leftarrow X_1 \\
0.6575 & \leftarrow X_2 \\
1.4155 & \leftarrow X_3 \\
-0.2043 & \leftarrow X_4 \\
-1.6054 & \leftarrow X_5 \\
1.1446 & \leftarrow X_6 \\
1.1033 & \leftarrow X_7 \\
0.7418 & \leftarrow X_8 \\
0.5035 & \leftarrow X_9 \\
1.5477 & \leftarrow X_{10}
\end{bmatrix}
\]

\[
\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
\hat{\sigma}^2_X = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu}_X)^2
\]

- \( \{X_1, X_2, \ldots, X_n\} \) is a **random sample** of size \( n \), i.e. a sequence of **independent and identically distributed (i.i.d)** random variables drawn from a distribution.

- \( x_i \) is an **observation** (realisation) of the random variable \( X_i \), i.e. the actual value assumed by that random variable in a given trial.

- The **sample mean** \( \hat{\mu}_X \) and the **sample variance** \( \hat{\sigma}^2_X \) are **unbiased estimators** of the **population mean** \( \mu_X \) and **population variance** \( \sigma^2_X \).
A boxplot is a graphical tool to summarise a distribution of data.

The box shows the first quartile (Q1), the second quartile (Q2, the median) with a line inside the box, and the third quartile (Q3).

The box represents the Inter-Quartile Range (IQR), i.e. the difference Q3-Q1.

The extension of the whiskers represents 1.5·IQR.

They roughly cover ~99% of the data, assuming a normal distribution.

Any data outside the whiskers is considered an outlier.

The Problem: Are They Different?
Statistical Hypothesis Testing

- **Statistical hypothesis testing** provides us with a mathematical framework to conduct **statistical inference** from the data.

- It compares the so-called **null hypothesis** $H_0$ against an **alternative hypothesis** $H_1$ or $H_A$.

- The comparison is **statistically significant** if the data are unlikely to be a realisation of the null hypothesis with respect to a chosen threshold, called **significance level** $\alpha$. In this case we **reject** the null hypothesis; in the opposite case, we **fail to reject** the null hypothesis.

---

Formulating the Problem

\[
\begin{align*}
H_0 : \mu_X &= \mu_Y \\
H_1 : \mu_X &\neq \mu_Y
\end{align*}
\]

\[
x = \begin{bmatrix}
3.1929 \\
0.6575 \\
1.4155 \\
-0.2043 \\
-1.6054 \\
1.1446 \\
1.1033 \\
0.7418 \\
0.5035 \\
1.5477
\end{bmatrix}
\]

\[
y = \begin{bmatrix}
0.7471 \\
1.6357 \\
1.0390 \\
1.1421 \\
1.5731 \\
3.5389 \\
1.9563 \\
0.6754 \\
2.6305 \\
1.9505
\end{bmatrix}
\]

\[
\hat{\mu}_X = 0.8497 \\
\hat{\mu}_Y = 1.6889
\]

Sample Mean
Test Statistic

Test statistic distribution $T$ under the null hypothesis $H_0$

$$\mathbb{P}[T \leq -t_{crit}|H_0] = \frac{\alpha}{2}$$

$$\mathbb{P}[T \leq -t_{stat}|H_0] = \frac{p}{2}$$

$$\mathbb{P}[T \geq t_{crit}|H_0] = \frac{\alpha}{2}$$

$$\mathbb{P}[T \geq t_{stat}|H_0] = \frac{p}{2}$$

$$1 - \alpha = \mathbb{P}[-t_{crit} < T < t_{crit}|H_0]$$
Test Statistic

Test statistic distribution $T$ under the null hypothesis $H_0$

\[ \mathbb{P}[T \leq -t_{\text{crit}}|H_0] = \frac{\alpha}{2} \]

\[ \mathbb{P}[T \leq -t_{\text{stat}}|H_0] = \frac{p}{2} \]

\[ \mathbb{P}[T \geq t_{\text{crit}}|H_0] = \frac{\alpha}{2} = \mathbb{P}[T \geq t_{\text{crit}}|H_0] \]

\[ \mathbb{P}[T \geq t_{\text{stat}}|H_0] = \frac{p}{2} = \mathbb{P}[T \geq t_{\text{stat}}|H_0] \]
# Types of Error

<table>
<thead>
<tr>
<th></th>
<th>We fail to reject $H_0$ [not statistically significant]</th>
<th>We reject $H_0$ [statistically significant]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true [e.g. systems are equivalent]</td>
<td>Correct conclusion [true negative]</td>
<td><strong>Type I error</strong> [false positive]</td>
</tr>
<tr>
<td></td>
<td>Probability $1 - \alpha$</td>
<td>Probability $\alpha$</td>
</tr>
<tr>
<td>$H_0$ is false [e.g. systems are not equivalent]</td>
<td><strong>Type II Error</strong> [false negative]</td>
<td><strong>Power</strong> (Correct conclusion) [true positive]</td>
</tr>
<tr>
<td></td>
<td>Probability $\beta$</td>
<td>Probability $1 - \beta$</td>
</tr>
</tbody>
</table>
Multiple Comparisons

- Type I errors concern the comparison of 2 samples
  - What happens when you need to compare \( c \) samples?
    - They originate \( m = \binom{c}{2} \) possible pairs to be compared, i.e. hypotheses to be tested simultaneously

- Performing multiple comparisons increases the Type I error probability, i.e. it is easier the reject the null hypothesis when you should not, since the \( m \) pairwise comparisons are independent

\[
\mathbb{P}(\text{No Type I Error}) = 1 - \alpha
\]

\[
\mathbb{P}(\text{No Type I Errors}) = \prod_{i=1}^{m} (1 - \alpha) = (1 - \alpha)^m
\]

\[
\mathbb{P}(\text{At Least One Type I Error}) = 1 - (1 - \alpha)^m
\]
General Linear Models (GLM)

Data = Model + Error

- A GLM explains the variation of a dependent variable (Data) in terms of a controlled variation of independent variables (Model) in addition to a residual uncontrolled variation (Error)

- Regression

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]

- ANalysis Of VAriance (ANOVA)

\[ y_{ij} = \mu_\cdot + \alpha_j + \varepsilon_{ij} \]

- the above regression model corresponds to the ANOVA one once you add as many \( x_{ij} \) predictors as many levels there are in the experimental condition \( \alpha_j \), e.g., by using dummy coding
Modelling System Effects (one-way ANOVA)

**ANOVA**

\[ y_{ij} = \mu_{..} + \alpha_j + \varepsilon_{ij} \]

- **Model**
  - \( \alpha_1 \)
  - \( \alpha_2 \)
  - \( \ldots \)
  - \( \alpha_q \)

- **Error**

<table>
<thead>
<tr>
<th>( \tau_1 )</th>
<th>( y_{11} )</th>
<th>( y_{12} )</th>
<th>( \ldots )</th>
<th>( y_{1q} )</th>
<th>( \mu_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau_2 )</td>
<td>( y_{21} )</td>
<td>( y_{22} )</td>
<td>( \ldots )</td>
<td>( y_{2q} )</td>
<td>( \mu_2 )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>( y_{ij} )</td>
<td>( \vdots )</td>
<td>( \mu_i )</td>
</tr>
<tr>
<td>( \tau_p )</td>
<td>( y_{p1} )</td>
<td>( y_{p2} )</td>
<td>( \ldots )</td>
<td>( y_{pq} )</td>
<td>( \mu_p )</td>
</tr>
</tbody>
</table>

| \( \mu_1 \) | \( \mu_2 \) | \( \mu_j \) | \( \mu_q \) | \( \mu_{..} \) |
Modelling System Effects (one-way ANOVA)

<table>
<thead>
<tr>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>( \ldots )</th>
<th>( \alpha_q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau_1 )</td>
<td>( y_{11} )</td>
<td>( y_{12} )</td>
<td>( y_{1q} )</td>
</tr>
<tr>
<td>( \tau_2 )</td>
<td>( y_{21} )</td>
<td>( y_{22} )</td>
<td>( y_{2q} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( y_{ij} )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( \tau_p )</td>
<td>( y_{p1} )</td>
<td>( y_{p2} )</td>
<td>( y_{pq} )</td>
</tr>
<tr>
<td>( \mu_1 )</td>
<td>( \mu_2 )</td>
<td>( \mu_j )</td>
<td>( \mu_q )</td>
</tr>
</tbody>
</table>
Estimators

- Grand mean
  \[ \hat{\mu}_{..} = \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} y_{ij} \]

- Marginal mean of the j-th system
  \[ \hat{\mu}_{..} = \frac{1}{p} \sum_{i=1}^{p} y_{ij} \]
  \[ \hat{\alpha}_{..} = \hat{\mu}_{..} - \hat{\mu}_{..} \]

- Predicted score
  \[ \hat{y}_{ij} = \hat{\mu}_{..} + \hat{\alpha}_{..} = \hat{\mu}_{..} \]

- Prediction error
  \[ \hat{e}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu}_{..} \]
Estimators
Assessment: ANOVA

- **Analysis of Variance (ANOVA)** was developed by statistician and evolutionary biologist Ronald Fisher (1890-1962).

- It provides a statistical test of whether or not the means of several groups are equal.

  - $H_0$ is the **null hypothesis** that all the means are equal.

- It partitions the observed variance in a particular variable into components attributable to different sources of variation.
Assessment: Sum of Squares

\[ y_{ij} - \mu.. = \mu.. - \mu.. + y_{ij} - \mu.. \]

- **Total Effect** \( \mu.. \)
- **System Effect** \( \alpha_j \)
- **Error Effect** \( \hat{\varepsilon}_{ij} \)

### Sum of squares (SS)

\[
SS_{\text{fact}} = \sum_{i=1}^{p} \sum_{j=1}^{q} \left( [\text{total} \mid \text{system} \mid \text{error}] \text{ effect} \right)^2
\]

### Variance break-down

\[
SS_{\text{Total}} = SS_{\text{System}} + SS_{\text{Error}}
\]
Assessment: Sum of Squares

- $SS_{\text{Error}}$ (Residual variation)
- $SS_{\text{System}}$ (Explained variation)
- $SS_{\text{Total}}$ (Total variation)
Assessment: Degrees of Freedom and Mean Squares

Degrees of Freedom (DF)

\[ DF_{\text{Total}} = pq - 1 \]

\[ DF_{\text{System}} = q - 1 \]

\[ DF_{\text{Error}} = (pq - 1) - (q - 1) = q(p - 1) \]

Mean Squares (MS)

\[ MS_{\text{fact}} = \frac{SS_{\text{fact}}}{DF_{\text{fact}}} \]

Assessment: F-test

- Let us assume to have a random sample of size \( n = pq \) (independence) from \( q \) normally-distributed random variables with same variance (homoskedasticity)

\[
Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)
\]

- Under the null hypothesis

\[
H_0 : \mu_1 = \mu_2 = \ldots = \mu_q
\]

the tests statistic is

\[
F_{stat} = \frac{MS_{System}}{MS_{Error}} \sim F\left(DF_{System}, DF_{Error}\right)
\]

\[
\alpha = \mathbb{P}[F \geq F_{crit}|H_0]
\]

\[
p = \mathbb{P}[F \geq F_{stat}|H_0]
\]
Assessment: F-test under $H_0$

\[ SS_{\text{Error}} \] (Residual variation)

\[ SS_{\text{System}} \] (Explained variation)

\[ SS_{\text{Total}} \] (Total variation)
Assessment: F-test under H₀

\[ \hat{\mu} \]

\( \mu \)

\( \hat{\mu}_1 \)

\( \hat{\mu}_2 \)

\( \hat{\mu}_3 \)

\( \hat{\mu}_{..} \)

\[ S S_{\text{Error}} \]

(Residual variation)

\[ S S_{\text{Total}} \]

(Total variation)
The Tukey Honestly Significant Difference (HSD) test creates confidence intervals $|t|$ for all pairwise differences between factor levels, while controlling the family error rate.

$$|t| = \left| \hat{\mu}_u - \hat{\mu}_v \right| > \frac{1}{\sqrt{2}} q_{\alpha, q, q(p-1)} \sqrt{\frac{2MS_{\text{Error}}}{p}}$$

- $\hat{\mu}_u$ and $\hat{\mu}_v$ are the marginal means of the two factor levels, i.e. the two systems to be compared.
- $q_{\alpha, q, q(p-1)}$ is the upper $100 \times (1 - \alpha)$-th percentile of the studentized range distribution, i.e. the distribution of the range of samples drawn from a normal distribution, considering $q$ systems to compare using $p$ topics.

Example: Analysis of TREC 8 Ad Hoc Systems (ordered by descending MAP)

Average Precision

Systems 1 to 65

Systems 66 to 129

Systems (ordered by descending MAP)
Example: Analysis of TREC 8 Ad Hoc
Example: Analysis of TREC 8 Ad Hoc

\[ y_{ij} = \mu + \tau_i + \alpha_j + \varepsilon_{ij} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
<th>( \hat{\omega}^2_{(f'act)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>167.9974</td>
<td>49</td>
<td>3.4285</td>
<td>251.9463</td>
<td>0</td>
<td>0.6559</td>
</tr>
<tr>
<td>System</td>
<td>60.0299</td>
<td>128</td>
<td>0.4690</td>
<td>34.4635</td>
<td>0</td>
<td>0.3991</td>
</tr>
<tr>
<td>Error</td>
<td>85.3502</td>
<td>6272</td>
<td>0.0136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>313.3375</td>
<td>6449</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Topic is a statistically significant and large-size effect.
- IR system is a statistically significant and large(medium)-size effect, quite smaller than topic.
Example: Analysis of TREC 8 Ad Hoc

- 129 systems amount to 8,256 system pairs to be compared
- According to Tukey HSD Test 3,425 system pairs are significantly different
- The top group consists of 7 systems
ANOVA Assumptions?

- **Independence**: topics and systems can be considered (reasonably) independent
- **Normality**: typical IR measures are bounded in [0, 1] so they cannot be normal
  - the normal distribution is unbounded
- **Homoskedasticity**: variance changes across systems

ANOVA is considered robust to violations of normality and also homoskedasticity when the sample sizes are equal and large (our typical case)

ANOVA Assumptions?

Real IR data

ANOVA assumptions
Questions?

**CAN MY BOYFRIEND COME ALONG?**

**I'M NOT YOUR BOYFRIEND!**

**YOU TOTALLY ARE!**

**I'M CASUALLY DATING A NUMBER OF PEOPLE.**

**BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.**

**YOUR MATH IS IRREFUTABLE.**

**FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.**
Foundations of Measurement
The Problem
Deeply Rooted...

IR

Experimentation
A Science of Measures...
User Models
A Science of Measures…

User Models

Top-heaviness
User Models

Top-heaviness

Incomplete Information
A Science of Measures…

User Models

Top-heaviness

Incomplete Information

Sensitivity
... But, Wait, Assumptions?

The operations you are allowed to perform with the values of a measure depend on the notion of measurement scale.

- mean
- variance
- correlation
- statistical tests
- ...
The operations you are allowed to perform with the values of a measure depend on the notion of measurement scale:

- mean
- variance
- correlation
- statistical tests

Do IR measures comply with those assumptions?

How much are (statistical) analyses impacted by departures from those assumptions?

What is the validity of our experiments?
… But, Wait, Assumptions?

The operations you are allowed to perform with the values of a measure depend on the notion of measurement scale:

- mean
- variance
- correlation
- statistical tests

Do IR measures comply with those assumptions?

How much are (statistical) analyses impacted by departures from those assumptions?

What is the validity of our experiments?
Measurement Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Basic Empirical Operations</th>
<th>Mathematical Group Structure</th>
<th>Permissible Statistics (invariantive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOMINAL</td>
<td>Determination of equality</td>
<td>Permutation group $x' = f(x)$</td>
<td>Number of cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any one-to-one substitution</td>
<td>Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Contingency correlation</td>
</tr>
<tr>
<td>ORDINAL</td>
<td>Determination of greater or less</td>
<td>Isotonic group $x' = f(x)$</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any monotonic increasing function</td>
<td>Percentiles</td>
</tr>
<tr>
<td>INTERVAL</td>
<td>Determination of equality of intervals or differences</td>
<td>General linear group $x' = ax + b$</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rank-order correlation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Product-moment correlation</td>
</tr>
<tr>
<td>RATIO</td>
<td>Determination of equality of ratios</td>
<td>Similarity group $x' = ax$</td>
<td>Coefficient of variation</td>
</tr>
</tbody>
</table>

An interval scale is an equi-spaced scale where a difference of one unit has the same meaning all over the range.

Temperature: An Interval Scale

- 20 °C is not twice as hot as 10 °C, i.e. multiplication and division are not allowed.
  - Division is **not invariant** wrt the transformation.

\[
\frac{20 \, ^\circ C}{10 \, ^\circ C} = 2 \quad \text{but} \quad \frac{68 \, ^\circ F}{50 \, ^\circ F} = 1.36
\]

- The increase between 10 °C and 20 °C is the same as the increase between 20 °C and 30 °C, i.e. addition and subtractions are allowed.
  - Subtraction is **invariant** wrt the transformation.

\[
\left\{
\begin{array}{l}
30 \, ^\circ C - 20 \, ^\circ C = 20 \, ^\circ C - 10 \, ^\circ C = 10 \, ^\circ C \\
86 \, ^\circ F - 68 \, ^\circ F = 68 \, ^\circ F - 50 \, ^\circ F = 18 \, ^\circ F
\end{array}
\right.
\]

- The ratio of intervals is **invariant** wrt to the transformation.

\[
\frac{20 \, ^\circ C - 10 \, ^\circ C}{30 \, ^\circ C - 20 \, ^\circ C} = 1 \quad \text{and} \quad \frac{68 \, ^\circ F - 50 \, ^\circ F}{86 \, ^\circ F - 68 \, ^\circ F} = 1
\]

\[F = \frac{9}{5}C + 32\]
Meaningfulness

- Statistical operations on measurements of a given scale are not appropriate or inappropriate per se but only relative to the kinds of statements made about them.

- The criterion of appropriateness for a statement about a statistical operation is that the statement be empirically meaningful in the sense that its truth or falsity must be invariant under permissible transformations of the underlying scale.

- Meaningfulness is a distinct concept from the one of truth of a statement and it is somehow close to the notion of invariance in geometry.

Temperature: Meaningfulness

\[ T^C_P = \begin{bmatrix} 2 & 2 & 4 & 8 & 36 \end{bmatrix} \]
\[ T^F_P = \begin{bmatrix} 35.6 & 35.6 & 39.2 & 46.4 & 96.8 \end{bmatrix} \]
\[ T^C_R = \begin{bmatrix} 1 & 2 & 4 & 15 & 34 \end{bmatrix} \]
\[ T^F_R = \begin{bmatrix} 33.8 & 35.6 & 39.2 & 59.0 & 93.2 \end{bmatrix} \]

- “The median temperature in Paris is the same as in Rome” is **meaningful**, since 4 = 4 in Celsius degrees and 39.2 = 39.2 in Fahrenheit degrees
  - interval scales are also ordinal and quantiles are an allowable operation on ordinal scales
- “The mean temperature in Paris is less than in Rome” is **meaningful** as well, since 10.4 < 11.2 in Celsius degrees and 50.72 < 52.16 in Fahrenheit degrees
  - addition and subtraction are allowable operations on an interval scale and, as a consequence, mean is invariant to affine transformations
- “The geometric mean of temperature in Paris is greater than in Rome” is **not meaningful**, since 5.40 > 5.27 in Celsius degrees and 46.74 < 48.17 in Fahrenheit degrees
  - geometric mean involves the multiplication and division of values, which is not a permitted operation on an interval scale
How Do IR Measures Look Like?

(a) Precision (and Recall).

(b) AP.

(c) RR.
How Do IR Measures Look Like?

(d) RBP, $p = 0.3$.  
(e) RBP, $p = 0.5$.  
(f) RBP, $p = 0.8$.  
(g) DCG, log base 2.
We may be tempted to compare the results of the arithmetic mean with those of the geometric mean to gain “more insights.”

We might observe that the arithmetic mean in Paris is less than in Rome – 10.4 < 11.2 in Celsius degrees – but the opposite is true when we consider the geometric mean – 5.40 > 5.27 in Celsius degrees.

We might thus highlight that this due to the fact that the first (and lowest) value 2 in Paris is double than 1 in Rome and that the geometric mean rewards gains at lowest values.

On the other hand, the arithmetic mean rewards gains at higher values and thus 8 in Paris is (almost) half than 15 in Rome and it contributes less.

However, if we consider exactly the same temperatures just on the Fahrenheit scale, we would reach opposite conclusions.
Overview of the TREC 2004 Robust Retrieval Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The robust retrieval track explores methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The retrieval task in the track is a traditional ad hoc retrieval task where the evaluation methodology emphasizes a system’s least effective topics. The most promising approach to improving poorly performing topics is exploiting text collections other than the target collection such as the web.

The 2004 edition of the track used 250 topics and required systems to rank the topics by predicted difficulty. The 250 topics within the test set allowed the stability of evaluation measures that emphasize poorly performing topics to be investigated. A new measure, a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results, shows promise of giving appropriate emphasis to poorly performing topics while being more stable at equal topic set sizes.

The ability to return at least passable results for any topic is an important feature of an operational retrieval system. While system effectiveness is generally reported as average effectiveness, an individual user does not see the average performance of the system, but only the effectiveness of the system on his or her requests. A user whose request retrieves nothing of interest is unlikely to be consoled by the fact that the system responds better to other people’s requests.

The TREC robust retrieval track was started in TREC 2003 to investigate methods for improving the consistency of retrieval technology. The first year of the track had two main technical results:

1. The track provided ample evidence that optimizing average effectiveness using the standard Cranfield methodology and standard evaluation measures further improves the effectiveness of the already-effective topics, sometimes at the expense of the poor performers.

2. The track results demonstrated that measuring poor performance is intrinsically difficult because there is so little signal in the sea of noise for a poorly performing topic. Two new measures devised to emphasize poor performers did so, but because there is so little information the measures are unstable. Having confidence in the conclusion that one system is better than another using these measures requires larger differences in scores than are generally observed in practice when using 50 topics.

The retrieval task in the track is a traditional ad hoc task. In addition to calculating scores using trec eval, each run is also evaluated using the two measures introduced in the TREC 2003 track that focus more specifically on the least-well-performing topics. The TREC 2004 track differed from the initial track in two important ways. First, the test set of topics consisted of 249 topics, up from 100 topics. Second, systems were required to rank the topics by predicted difficulty, with the goal of eventually being able to use such predictions to do topic-specific processing.

This paper presents an overview of the results of the track. The first section describes the data used in the track, and the following section gives the retrieval results. Section 3 investigates how accurately systems can predict which topics are difficult. Since one of the main results of the TREC 2003 edition of the track was that the poor performance is hard to measure with 50 topics, section 4 examines the stability of the evaluation measures for larger topic set sizes. The final section looks at the future of the track.

1 The Robust Retrieval Task

As mentioned, the task within the robust retrieval track is a traditional ad hoc task. Since the TREC 2003 track had shown that 50 topics was not sufficient for a stable evaluation of poorly performing topics, the TREC 2004 track used...
Measurement Issues in IR?

- Comparing system performance
- Topic difficulty and robust retrieval
- Score transformation and standardization techniques
- ...

- Statistical significance testing
  - Sign Test: ordinal scale
  - Wilcoxon Rank Sum Test: ordinal scale
  - Wilcoxon Signed Rank Test: interval scale
  - Student’s t Test: interval scale
  - ANOVA: interval scale
  - Kruskal-Wallis Test: ordinal scale
  - Friedman Test: ordinal scale


The Plan
There exists an **empirical relation** \( \succ \) which orders entities on the basis of their attributes.

- E.g., you can compare two rods and determine which is longer or whether they are equal.
- The empirical relation may support **concatenation** \( \circ \)
  - E.g., the concatenation of a rod with another one is longer than both of them.

There exists an **homomorphism** \( M \), the **measurement scale**, which maps entities into numbers and the empirical relation into a numerical relation which preserves the ordering (and concatenation).

\[
(J(E, \succ, \circ) \xrightarrow{M} (\mathbb{R}, >, +))
\]

\[
e_1 \succ e_2 \implies M(e_1) > M(e_2)
\]

\[
M(e_1 \circ e_2) = M(e_1) + M(e_2)
\]


From Theory to Trees...

Real World

- Papaya Tree
  - taller than Banana Tree
  - taller than Ananas Tree
- Banana Tree
  - taller than Ananas Tree
- Papaya Tree much taller than Ananas Tree
- Banana Tree much taller than Ananas Tree

Empirical Relations

Number System

- \( M(\cdot) \)
- \( \mathbb{R}_0^+ \)
- \( M(\text{Papaya Tree}) > M(\text{Banana Tree}) \)
- \( M(\text{Papaya Tree}) > M(\text{Ananas Tree}) \)
- \( M(\text{Banana Tree}) > M(\text{Ananas Tree}) \)
- \( M(\text{Papaya Tree}) > M(\text{Ananas Tree}) + 50 \)
- \( M(\text{Banana Tree}) > M(\text{Ananas Tree}) + 50 \)

Mathematical Relations

\( M_R(\cdot) \)
“In the physical sciences there is usually an empirical ordering of the quantities we wish to measure [...] Such a situation does not hold for information retrieval. There is no empirical ordering for retrieval effectiveness and therefore any measure of retrieval effectiveness will by necessity be artificial”

C. J. “Keith” van Rijsbergen

Measurement Issues: Ordering?

More Measurement Issues: Intervals?

To determine whether a measure is interval-based

- We need to have a notion of interval among runs
- We need to have a notion of length of an interval among runs

Which runs fall in-between?

\[ \ell = 1, 2, 3, \ldots, ? \]
**Table 1**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Basic Empirical Operations</th>
<th>Mathematical Group Structure</th>
<th>Permissible Statistics (invariantive)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nominal</strong></td>
<td>Determination of equality</td>
<td><em>Permutation group</em> $x' = f(x)$</td>
<td>Number of cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any one-to-one substitution</td>
<td>Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Contingency correlation</td>
</tr>
<tr>
<td><strong>Ordinal</strong></td>
<td>Determination of greater or less</td>
<td><em>Isotonic group</em> $x' = f(x)$</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any monotonic increasing function</td>
<td>Percentiles</td>
</tr>
<tr>
<td><strong>Interval</strong></td>
<td>Determination of equality of intervals or differences</td>
<td><em>General linear group</em> $x' = ax + b$</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rank-order correlation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Product-moment correlation</td>
</tr>
<tr>
<td><strong>Ratio</strong></td>
<td>Determination of equality of ratios</td>
<td><em>Similarity group</em> $x' = ax$</td>
<td>Coefficient of variation</td>
</tr>
</tbody>
</table>
Measures: Don’t Give Up!

“Measure what is measurable and make measurable what is not”

Galileo Galilei (1564-1642)
A General Theory of IR Measures

Approach

- Define what orders and intervals among runs are
- Using orders and intervals of runs, introduce a proper structure among runs which allows us to define an interval scale measure by construction
- Try to transform IR measures to the interval scale one and determine whether they are interval scales not

<TL;DR>

- Set-based measure are interval scales
- Rank-based measures are interval scales only under very strict conditions, hardly met in practice
- When you go multi-graded… be though!

Set-based Measures: Findings

- Traditional set-based measures are all interval scales

- Precision

  \[ P(\hat{r}) = \frac{1}{N} \sum_{i=1}^{N} \hat{r}_i = \frac{1}{N} \text{SBTO}(\hat{r}) \]

- Recall

  \[ R(\hat{r}) = \frac{1}{RB} \sum_{i=1}^{N} \hat{r}_i = \frac{1}{RB} \text{SBTO}(\hat{r}) \]

- F-measure

  \[ F(\hat{r}) = 2 \frac{P(\hat{r}) \cdot R(\hat{r})}{P(\hat{r}) + P(\hat{r})} = \frac{2}{N + RB} \text{SBTO}(\hat{r}) \]
Rank-based Measures: Findings

- RBP with $p \leq \frac{1}{2}$ is ordinal

- Only RBP with $p = \frac{1}{2}$ is interval

\[
\text{RBP}_{\frac{1}{2}} = \frac{1}{2} \sum_{i=1}^{N} \frac{1}{2^{i-1}} \hat{r}[i] = \frac{1}{2^N} \sum_{i=1}^{N} 2^{N-i} \hat{r}[i] = \frac{1}{2^N} \text{RBTO} (\hat{r})
\]

- RBP with $p > \frac{1}{2}$ and AP are not even ordinal
Experiments
### Binary Relevance – T08, 8,256 system pairs compared

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Significantly Different Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kruskal-Wallis Test</td>
</tr>
<tr>
<td>Precision</td>
<td>1,566 (18.97%)</td>
</tr>
<tr>
<td>Recall</td>
<td>1,748 (21.17%)</td>
</tr>
<tr>
<td>F-measure</td>
<td>1,721 (20.85%)</td>
</tr>
<tr>
<td>SBTO</td>
<td>1,566 (18.97%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Significantly Different Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kruskal-Wallis Test</td>
</tr>
<tr>
<td>RBP $p = 1/2$</td>
<td>1,677 (20.31%)</td>
</tr>
<tr>
<td>RBP $p = 0.2$</td>
<td>1,675 (20.29%)</td>
</tr>
<tr>
<td>RBP $p = 0.8$</td>
<td>1,783 (21.60%)</td>
</tr>
<tr>
<td>AP</td>
<td>1,824 (22.09%)</td>
</tr>
<tr>
<td>RBTO</td>
<td>1,677 (20.31%)</td>
</tr>
</tbody>
</table>

### We Understood Everything, right? Significance Tests

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Significantly Different Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kruskal-Wallis Test</td>
</tr>
<tr>
<td><strong>Binary Relevance – T08, 8,256 system pairs compared</strong></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>1,566 (18.97%)</td>
</tr>
<tr>
<td>Recall</td>
<td>1,748 (21.17%)</td>
</tr>
<tr>
<td>F-measure</td>
<td>1,721 (20.85%)</td>
</tr>
<tr>
<td>SBTO</td>
<td>1,566 (18.97%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Significantly Different Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kruskal-Wallis Test</td>
</tr>
<tr>
<td><strong>Binary Relevance – T08, 8,256 system pairs compared</strong></td>
<td></td>
</tr>
<tr>
<td>RBP $p = 1/2$</td>
<td>1,677 (20.31%)</td>
</tr>
<tr>
<td>RBP $p = 0.2$</td>
<td>1,675 (20.29%)</td>
</tr>
<tr>
<td>RBP $p = 0.8$</td>
<td>1,783 (21.60%)</td>
</tr>
<tr>
<td>AP</td>
<td>1,824 (22.09%)</td>
</tr>
<tr>
<td>RBTO</td>
<td>1,677 (20.31%)</td>
</tr>
</tbody>
</table>

### Binary Relevance – T08, 8,256 system pairs compared

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Significantly Different Pairs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kruskal-Wallis Test</td>
<td>ANOVA</td>
</tr>
<tr>
<td>Precision</td>
<td>1,566 (18.97%)</td>
<td>2,785 (33.73%)</td>
</tr>
<tr>
<td>Recall</td>
<td>1,748 (21.17%)</td>
<td>3,259 (39.47%)</td>
</tr>
<tr>
<td>F-measure</td>
<td>1,721 (20.85%)</td>
<td>3,081 (37.32%)</td>
</tr>
<tr>
<td>SBTO</td>
<td>1,566 (18.97%)</td>
<td>2,785 (33.73%)</td>
</tr>
</tbody>
</table>

### Binary Relevance – T08, 8,256 system pairs compared

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Significantly Different Pairs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kruskal-Wallis Test</td>
<td>ANOVA</td>
</tr>
<tr>
<td>RBP $p = 1/2$</td>
<td>1,677 (20.31%)</td>
<td>2,861 (34.65%)</td>
</tr>
<tr>
<td>RBP $p = 0.2$</td>
<td>1,675 (20.29%)</td>
<td>2,198 (26.62%)</td>
</tr>
<tr>
<td>RBP $p = 0.8$</td>
<td>1,783 (21.60%)</td>
<td>3,476 (42.10%)</td>
</tr>
<tr>
<td>AP</td>
<td>1,824 (22.09%)</td>
<td>3,320 (40.21%)</td>
</tr>
<tr>
<td>RBTO</td>
<td>1,677 (20.31%)</td>
<td>2,861 (34.65%)</td>
</tr>
</tbody>
</table>

### We Understood Everything, right? Significance Tests

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Significantly Different Pairs</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary Relevance – T08, 8,256 system pairs compared</strong></td>
<td><strong>Kruskal-Wallis Test</strong></td>
<td><strong>ANOVA</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>1,566 (18.97%)</td>
<td>2,785 (33.73%)</td>
<td></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>1,748 (21.17%)</td>
<td>3,259 (39.47%)</td>
<td></td>
</tr>
<tr>
<td><strong>F-measure</strong></td>
<td>1,721 (20.85%)</td>
<td>3,081 (37.32%)</td>
<td></td>
</tr>
<tr>
<td><strong>SBTO</strong></td>
<td>1,566 (18.97%)</td>
<td>2,785 (33.73%)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Measure Pair</strong></th>
<th><strong>Significantly Different Pairs</strong></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary Relevance – T08, 8,256 system pairs compared</strong></td>
<td><strong>Kruskal-Wallis Test</strong></td>
<td><strong>ANOVA</strong></td>
<td></td>
</tr>
<tr>
<td><strong>RBP $p = 1/2$</strong></td>
<td>1,677 (20.31%)</td>
<td>2,861 (34.65%)</td>
<td></td>
</tr>
<tr>
<td><strong>RBP $p = 0.2$</strong></td>
<td>1,675 (20.29%)</td>
<td>2,198 (26.62%)</td>
<td></td>
</tr>
<tr>
<td><strong>RBP $p = 0.8$</strong></td>
<td>1,783 (21.60%)</td>
<td>3,476 (42.10%)</td>
<td></td>
</tr>
<tr>
<td><strong>AP</strong></td>
<td>1,824 (22.09%)</td>
<td>3,320 (40.21%)</td>
<td></td>
</tr>
<tr>
<td><strong>RBTO</strong></td>
<td>1,677 (20.31%)</td>
<td>2,861 (34.65%)</td>
<td></td>
</tr>
</tbody>
</table>

We Understood Everything, right? Kendall’s Tau

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision vs SBTO</td>
<td>0.9998</td>
</tr>
<tr>
<td>Recall vs SBTO</td>
<td>0.8591</td>
</tr>
<tr>
<td>F-measure vs SBTO</td>
<td>0.9670</td>
</tr>
<tr>
<td>Precision vs Recall</td>
<td>0.8588</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBP ( p = 1/2 ) vs RBTO</td>
<td>1.0000</td>
</tr>
<tr>
<td>RBP ( p = 0.2 ) vs RBTO</td>
<td>0.9225</td>
</tr>
<tr>
<td>RBP ( p = 0.8 ) vs RBTO</td>
<td>0.9043</td>
</tr>
<tr>
<td>AP vs RBTO</td>
<td>0.7439</td>
</tr>
</tbody>
</table>
## Binary Relevance – T08

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Topic-by-Topic</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision vs SBTO</td>
<td>1.0000</td>
<td>0.9998</td>
</tr>
<tr>
<td>Recall vs SBTO</td>
<td>1.0000</td>
<td>0.8591</td>
</tr>
<tr>
<td>F-measure vs SBTO</td>
<td>1.0000</td>
<td>0.9670</td>
</tr>
<tr>
<td>Precision vs Recall</td>
<td>1.0000</td>
<td>0.8588</td>
</tr>
</tbody>
</table>

## Binary Relevance – T08

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Topic-by-Topic</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBP $p = 1/2$ vs RBTO</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>RBP $p = 0.2$ vs RBTO</td>
<td>0.9985</td>
<td>0.9225</td>
</tr>
<tr>
<td>RBP $p = 0.8$ vs RBTO</td>
<td>0.8553</td>
<td>0.9043</td>
</tr>
<tr>
<td>AP vs RBTO</td>
<td>0.6099</td>
<td>0.7439</td>
</tr>
</tbody>
</table>
We Understood Everything, right? Kendall’s Tau

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Topic-by-Topic</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision vs SBTO</td>
<td>1.0000</td>
<td>0.9998</td>
</tr>
<tr>
<td>Recall vs SBTO</td>
<td>1.0000</td>
<td>0.8591</td>
</tr>
<tr>
<td>F-measure vs SBTO</td>
<td>1.0000</td>
<td>0.9670</td>
</tr>
<tr>
<td>Precision vs Recall</td>
<td>1.0000</td>
<td>0.8588</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure Pair</th>
<th>Topic-by-Topic</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBP $p = 1/2$ vs RBTO</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>RBP $p = 0.2$ vs RBTO</td>
<td>0.9985</td>
<td>0.9225</td>
</tr>
<tr>
<td>RBP $p = 0.8$ vs RBTO</td>
<td>0.8553</td>
<td>0.9043</td>
</tr>
<tr>
<td>AP vs RBTO</td>
<td>0.6099</td>
<td>0.7439</td>
</tr>
</tbody>
</table>
Worser and Worser Measurement Issues

- We started with a concern on whether IR measures should be **interval scales** or not
  - Whatever your stance on this, it matters since it changes the number of SSD pairs
- Independently from being interval scales or not, we found that **length** of the run and, especially, **recall base** change the scale you use from topic to topic
  - When you are aggregating across topics, e.g. **averaging**, you are mixing numbers from different scales
  - We can control the run length but not the recall base

Precision and Recall: An Intuition

\[ P(k) = \frac{1}{k} \sum_{n=1}^{k} r_n \]
\[ R(k) = \frac{1}{RB} \sum_{n=1}^{k} r_n \]

- Topics are like planets
- Precision is (somehow) like mass
- Recall is (somehow) like weight
- One the same planet, mass and weight order bodies in the same way
- However...
- You could average mass across planets but not weight
Lessons Learnt

- It is possible to develop a theory of IR measures grounded in the representational theory of measurement.
- We determined the scale properties of several state-of-the-art IR measures:
  - issues with intervalness but even more with recall base and run length.
- Experimental results agree with the expected properties of the measures:
  - you need to deep dive to really understand the behavior.
  - sizeable impact on both correlation analysis and statistical significance testing.
It is possible to develop a theory of IR measures grounded in the representational theory of measurement.

We determined the scale properties of several state-of-the-art IR measures, issues with intervalness but even more with recall base and run length.

Experimental results agree with the expected properties of the measures, you need to deep dive to really understand the behavior.

Sizeable impact on both correlation analysis and statistical significance testing.

✦ We need to **rethink** how we use our analytical tools and how we **explain** their outcomes.

✦ What is the **validity** of our experiments is still an open question.

✦ **Meaningfulness** should be a central concern in IR.
No Shortcuts
No Shortcuts
Questions?

IT TOOK WEEKS BUT I'VE CALCULATED A NEW THEORY ABOUT THE ORIGIN OF THE UNIVERSE.

ACCORDING TO MY CALCULATIONS IT DIDN'T START WITH A "BIG BANG" AT ALL - IT WAS MORE OF A "PHHBBWT" SOUND.

YOU MAY BE WONDERING ABOUT THE PRACTICAL APPLICATIONS OF THE "LITTLE PHHBBWT" THEORY.

I WAS WONDERING WHEN YOU'LL GO AWAY.
Questions?
Reproducibility
Reproducibility: Why?

What we find reported in papers…

P I X A R
Reproducibility: Why?

What we find reported in papers...

...what happens to us
Reproducibility: How?

Everybody likes reproducibility…
Reproducibility: How?

Everybody likes reproducibility…

…as soon as someone else does it
Reproducibility: What?

We know, reproducibly is at the core of science...

https://xkcd.com/242/
Reproducibility: What?

We know, reproducibly is at the core of science…

…but reproducing research is not new research

https://xkcd.com/242/
The “Reproducibility” Nautilus

Reproduce
Different Data/Setup
Same Task/Goal
Same/Different Software
Different Group

Replicate
Same Data/Setup
Same Task/Goal
Same/Different Software
Different Group

Repeat
Same Data/Setup
Same Task/Goal
Same Software
Same Group

Experiment

Generalize/Re-use
Different Data/Setup
Different Task/Goal
Same/Different Software
Different Group
The “Reproducibility” Nautilus

Is it really that easy?

DON'T WORRY. YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.

YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.

ARE THERE INSTRUCTIONS FOR HOW TO USE IT? I DOUBT IT.

IS THE CODE COMMENTED? NOT LIKELY.

WHERE ARE THE FILES? WHO KNOWS.

THIS IS GOING TO BE PAINFUL, ISN'T IT?

JUST A SCRATCH.

Different Group

Replicate
Same Data/Setup
Same Task/Goal
Same/Different Software
Different Group

Experiment

Repeat
Same Data/Setup
Same Task/Goal
Same Software
Same Group

Generalize/Re-use
Different Data/Setup
Different Task/Goal
Same/Different Software
Different Group

Is it really that easy?


http://phdcomics.com
The “Reproducibility” Nautilus

Cranfield/Evaluation Campaigns

...BUT...

- **format babele**, lack of data and metadata formats
- shared data and code **repositories**, difficulties in adoption (DIRECT, EvaluatIR, OpenRuns, TIRA, EaaS, …)
- scripts are not **workflows**, actionable papers, …

...BUT...

Are all of these core IR research? Cultural mismatch
The “Reproducibility” Nautilus

Somehow standard approach in IR evaluation

...BUT...

✦ typically done with-in group, in a repeatability-like fashion
✦ how to quantify when “reproduced” - same ranked list, correlation among ranked lists, same effectiveness score, confidence bounds on effectiveness score, close distributions of effectiveness score, similar statistical significance (p-values, effect sizes, ...), ...
✦ what about the user-side 😱?

...Initial investigation in CENTRE@CLEF/NTCIR/TREC...

http://www.centre-eval.org/
Largely unexplored: it means turning IR into a predictive science

...Some seeds...

- Fuhr’s Salton award talk
- query performance prediction
- performance modelling and break-down via GLMM, ANOVA
- ML for predicting best system configuration

...Manifesto from Dagstuhl Perspectives Workshop 17442...
The “Reproducibility” Nautilus

We need a shared conceptual framework

A starting point: the PRIMAD model developed during the Dagstuhl Seminar 16041…
What about Introducing Badges?

What about Introducing Badges?

- Better science
- Better baselines (and less effort)
- Improve community

---

What about Introducing Badges?

- Better science
- Issues for industry, proprietary/confidential data
- Slow-down new ideas
- Not trivial logistics
- Additional effort for authors and reviewers
- Publish or perish
- High-barrier for small groups

What about Introducing Badges?

- Additional effort for authors and reviewers
- Issues for industry, proprietary/confidential data
- Class A and B papers
- Not all research is reproducible

What about Introducing Badges?

- Issues for industry, proprietary/confidential data
- Additional effort for authors and reviewers
- Not all research is reproducible
- Trust
- Class A and B papers

What about Introducing Badges?

- Issues for industry, proprietary/confidential data
- Additional effort for authors and reviewers
- Slow-down new ideas

Would Badges Change Your Research?

![Bar chart showing frequency and cumulative percent responses to the question: Would Badges Change Your Research?](chart.png)

- **Definitely Yes**: Frequency 25, Cumulative Percent 25
- **Probably Yes**: Frequency 30, Cumulative Percent 55
- **Might or Might Not**: Frequency 35, Cumulative Percent 90
- **Probably No**: Frequency 40, Cumulative Percent 100
- **Definitely Not**: Frequency 50

The chart indicates a trend where the likelihood of badges changing research increases with responses from "Definitely Yes" to "Definitely Not."
Would Badges Change Your Research?

- Better science
- More care
- More effort
- Open datasets
Would Badges Change Your Research?

- Better science
- More care
- More effort
- Already do it
- Publish or perish

![Bar chart showing frequency and cumulative percent for different responses to the question: Would Badges Change Your Research?]

- Definitely Yes: Frequency 20, Cumulative Percent 50
- Probably Yes: Frequency 30, Cumulative Percent 80
- Might or Might Not: Frequency 40, Cumulative Percent 100
- Probably No: Frequency 40, Cumulative Percent 100
- Definetely Not: Frequency 30, Cumulative Percent 100
Would Badges Change Your Research?

- More care
- More effort
- Already do it
- Issues for industry
- Not an incentive

![Graph showing frequency and cumulative percent for responses to the question: Would Badges Change Your Research?](chart.png)
Would Badges Change Your Research?

- Already do it
- Issues for industry
- Not an incentive
Would Badges Change Your Research?

- Already do it
- Issues for industry
- Not an incentive
- Trust

Frequency

Cumulative Percent
Reproducibility: Some Needs

- Shift in culture
  - more work needed to put reproducibility in action
  - what about the pressure to publish?
  - acknowledgment in careers

- Systematic but focused approach
  - how to choose what to reproduce?

- Quantitative assessment
  - when do we consider something as “reproduced”? 

- Infrastructures (evaluation campaigns?)
  - lightweight tools and protocols… but they need adoption!
Questions?

---

**THE THESIS**
At a grad student "social"...

Hello, Nerd. I am Geekity.

WH...? You mean the Geekity that aced her quals her first year?

That was a long time ago. Listen to me. I know why you're here.

**UM... THE FREE FOOD?**

You are here because of the question: what is the Repro?

**B-But, I thought you weren't supposed to figure it out until your 5th year...**

That's just what they want you to believe. Follow me.

---