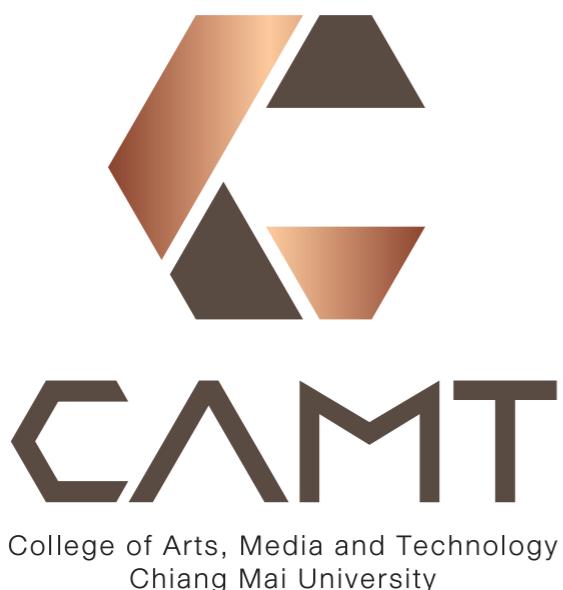


SE 481 Introduction to Information Retrieval

Module #5 — Evaluation



Passakorn Phannachitta, D.Eng.

passakorn.p@cmu.ac.th

College of Arts, Media and Technology
Chiang Mai University, Chiangmai, Thailand

Agenda

- Evaluation metrics in IR

It is all about telling whether users are happy

- Is the returned search products relevant to users?
- Do the users click the search results a lot?
- How much users spend a lot of money after searching for products using the search engine?
- How long does it take for a user to reach the solutions after starting the search

Happiness — too elusive to measure

- It does not mean it cannot be measured
 - Similar to unit test vs acceptance test
- The unit-level test in IR is the **degree of relevancy** or **relevance**

Measuring relevance

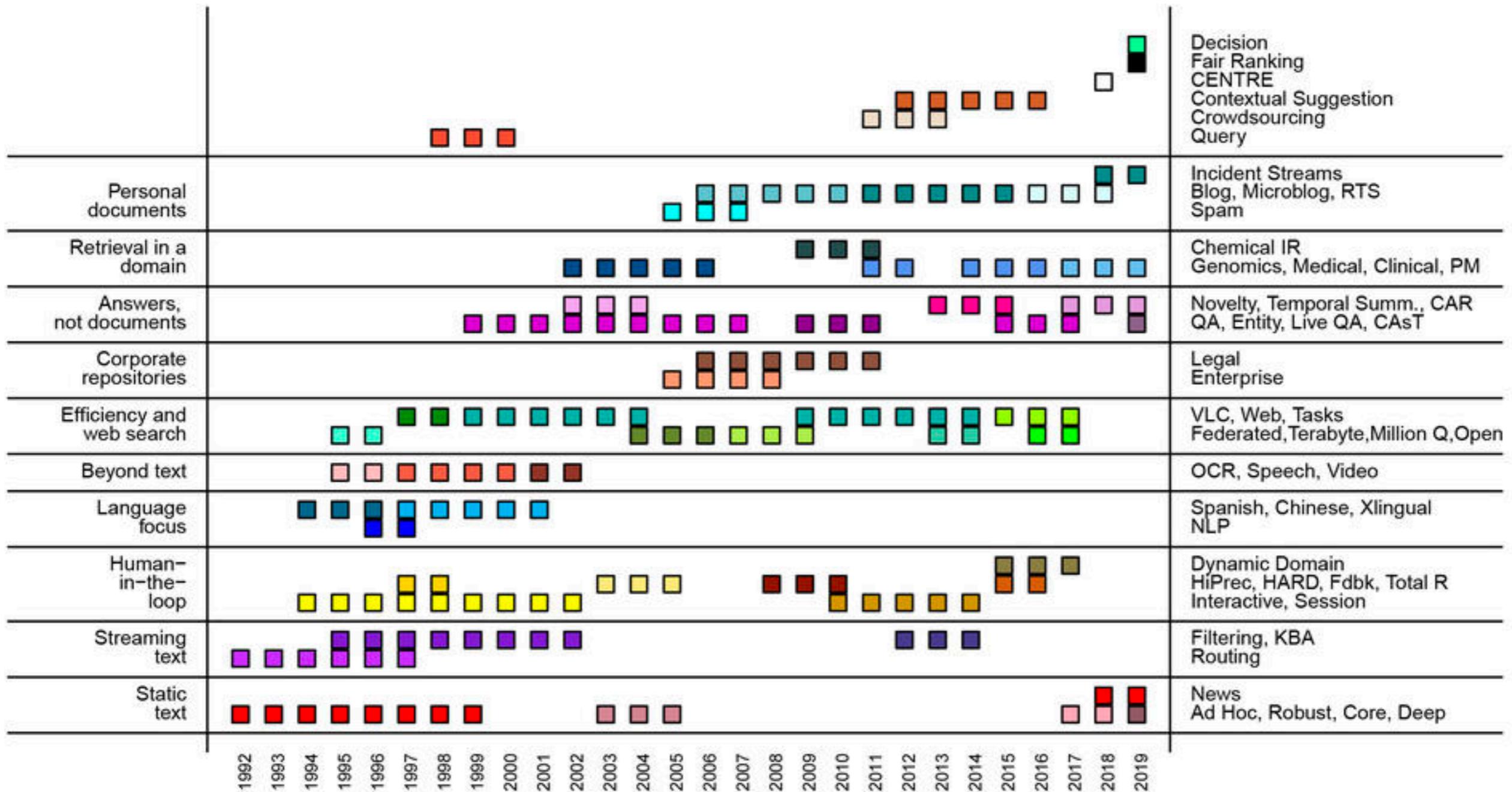
- A benchmark document collection
- A benchmark suite of queries
- An assessment of either **relevant** or **non-relevant** for each query and each document

Relevance judgement

- Issue?
 - Document sets are large, e.g., web pages in the internet
 - Rapid increasing rates of documents everyday
- Crowd-sourcing is commonly used
 - variance and quality might not be sufficiently high
- Testing on standard benchmarking collection is important
 - Difficult and representative questions

Standard benchmark

- Text REtrieval conference — TREC datasets
 - Series of workshops focusing on a list of different information retrieval (IR) research areas
 - Start in 1992 (~30 years)



Ref: <https://www.nist.gov/image/tracksjpg>

- Some recent interesting tracks



TREC 2021 Deep Learning Track Data Refresh

New, larger, cleaner corpus
Document dataset 3.7 times larger
Passage dataset 15.6 times larger
4x more passages per document
UTF-8. JSONL. Eliminating problems with whitespace and character sets

Realistic scenario
Start with documents, and generate candidate passages
Passage↔Document mapping is available and can be used in modeling
Our previous passage dataset was selected in a query-biased fashion, so using a Passage↔Document mapping would leak ground truth information

Basis for future leaderboards and tasks
Updated MS MARCO leaderboards
Support for docker and shared community resources
More metadata, updated ORCAS click data

Text REtrieval Conference (TREC)
...to encourage research in information retrieval from large text collections.

MS MARCO

Evaluating an IR system

- User's need is translated into a query prior to any assessment
- Relevance is assessed relative to the user need, **not the query**
- E.g.,
 - Information need: I want to extract all the textual information from a pdf file using java
 - A possible query: **get text java pdf**
 - A **maybe** better query: **pdf parsing java**
 - **What to access:** Whether the returned documents can help the user to find a solution, not whether they contain these words.

Standard evaluation metric family

- Precision (for ranking)
- Recall (for ranking)
- DCG

Precision and Recall

- In unranked binary assessment they are defined as—
 - Precision — ratio between the correct prediction and total case of predicting true.
 - Recall — ratio between the correct prediction and total correct case.
- Let's discuss some terms first
 - True Positives (TP) — number of predicted positive in the actual positive group
 - True Negatives (TN) — number of predicted negative in the actual negative group
 - False Positives (FP) — number of predicted positive in the actual negative group
 - False Negatives (FN) — number of predicted negative in the actual positive group

Precision and Recall in unranked problems

- Precision = $\frac{TP}{TP + FP}$
- Recall = $\frac{TP}{TP + FN}$
- Example

	Predicted Positive	Predicted Negative
Actual Positive	10 (TP)	15 (FN)
Actual Negative	25 (FP)	100 (TN)

- Precision = $(10) / (10 + 25) = 0.29$
- Recall = $(10) / (10 + 15) = 0.4$

Standard evaluation metrics for ranking

- Binary relevance
 - Precision@K (P@K)
 - Recall@K (R@K)
 - Mean average precision (mAP)
 - Mean reciprocal rank (mRR)
- Multiple levels of relevance
 - Normalized discounted cumulative gain (NDCG)

Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K
- E.g., Rank #1

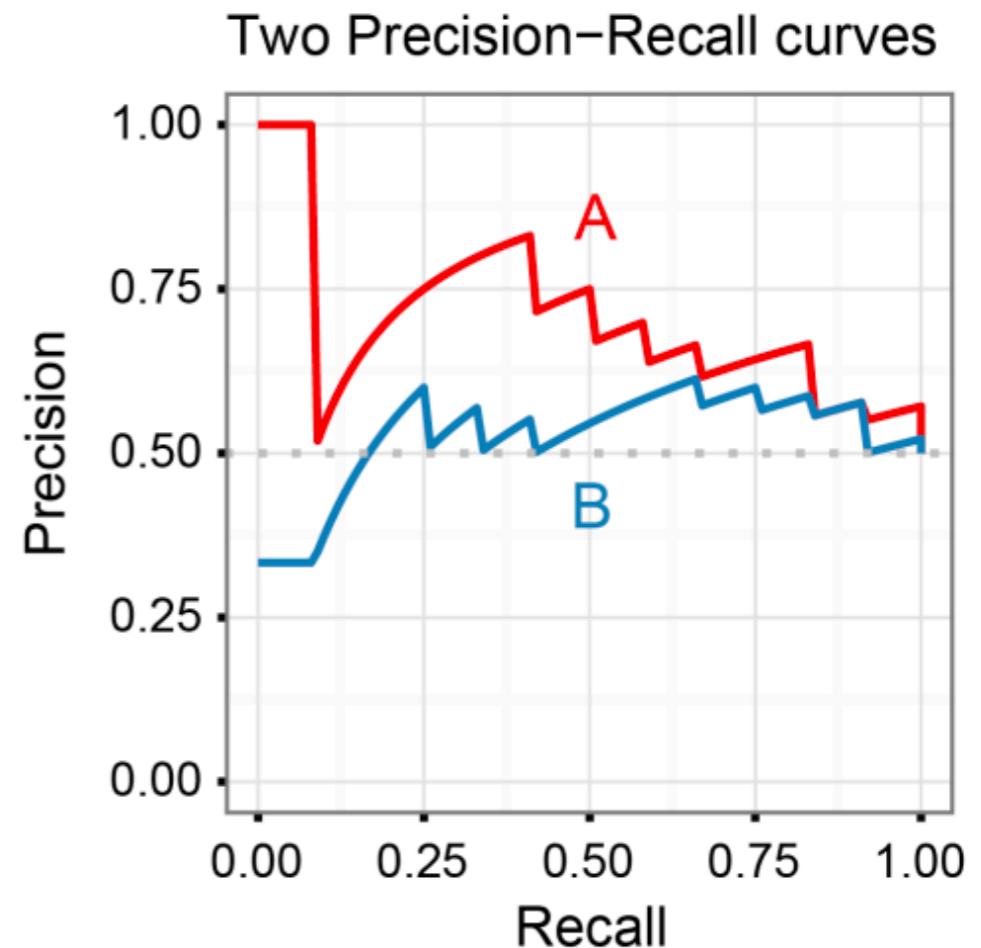


- Precision@3 = 1/3
- Precision@4 = 2/4
- Precision@5 = 3/5

- Recall@K is in similar fashion

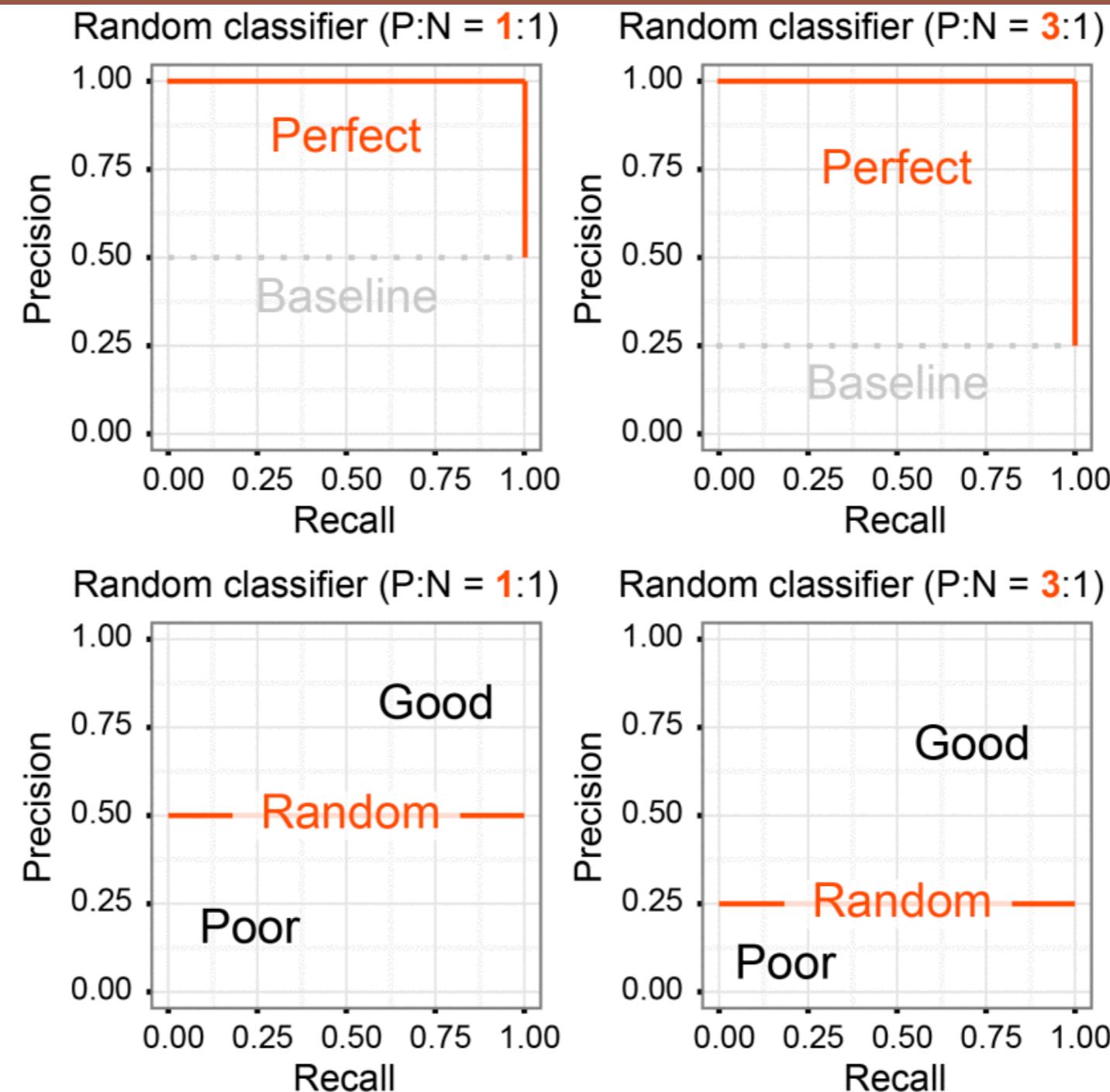
Precision-recall curve

- For comparing two or more information retrieval systems



Ref — <https://stackoverflow.com/questions/40865645/confusion-about-precision-recall-curve-and-average-precision>

Precision-recall curve



Mean average precision

- AP — the average of the precision scores at the rank locations of each relevant document
- mAP — the mean of the average precision scores for a group of queries
 - Macro average — All queries are considered equal
- If a relevant document never gets retrieved, we assume its corresponding precision as zero.

Average precision



There are 5 documents relevant to query #1

Ranking for query #1



	Recall	Precision
0.2	0.2	1.0
0.4	0.4	0.5
0.4	0.4	0.67
0.6	0.6	0.5
0.6	0.6	0.43
0.8	0.8	0.38
1.0	1.0	0.44

$$\text{Average precision for query } \#1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5) / 5 = 0.62$$



There are 3 documents relevant to query #2

Ranking for query #2



	Recall	Precision
0.0	0.0	0.0
0.33	0.33	0.5
0.33	0.33	0.33
0.33	0.33	0.25
0.67	0.67	0.4
0.67	0.67	0.33
1.0	1.0	0.43
1.0	1.0	0.38
1.0	1.0	0.33
1.0	1.0	0.3

$$\text{Average precision for query } \#2 = (0.5 + 0.4 + 0.43) / 3 = 0.44$$

Mean average precision



Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

Average precision for query #1 = $(1.0 + 0.67 + 0.5 + 0.44 + 0.5) / 5 = 0.62$



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

Average precision for query #2 = $(0.5 + 0.4 + 0.43) / 3 = 0.44$

Mean average precision = $(0.62 + 0.44) / 2 = 0.53$

Mean reciprocal rank

- RR — reciprocal of the rank at which first correct response returned (or 0 if none)
- mRR — the mean of the RR scores for a group of queries
- Only cares about the single highest-ranked relevant item.

Reciprocal rank

Ranking for query #1



Reciprocal Rank = 1/5

Ranking for query #2



Reciprocal Rank = 1/2

Ranking for query #3



Reciprocal Rank = 1/1

Mean reciprocal rank

Ranking for query #1



Reciprocal Rank = 1/5

Ranking for query #2



Reciprocal Rank = 1/2

Ranking for query #3



Reciprocal Rank = 1/1

Mean reciprocal rank = $(1/5 + 1/2 + 1/1) / 3 = 1.7$

Beyond binary relevance

get text pdf java

All Images Videos News Maps Settings

Thailand (en) Safe search: strict Any time

PDF Java Library - Comperhensive Tutorials AD
e-iceblue.com | Report Ad
Create, Process, Save or Convert PDF Documents in Java-based Applications. Convert PDF to PDF/A, Word, HTML, Image, etc.

How to get raw text from pdf file using java - Stack Overflow
https://stackoverflow.com/questions/18098400/how-to-get-raw-text-from-pdf-file-usin...
I have some pdf files, Using pdfbox i have converted them into text and stored into text files, Now from the text files i want to remove. Hyperlinks; All special characters; Blank lines; headers footers of pdf files "1","2", "a", "bullets", etc. I want to get valid text line by line like this:

Search and Get Text from Pages of PDF Document Java ...
https://docs.aspose.com/pdf/java/search-and-get-text-from-pdf/
This article explains how to use various tools to search and get a text from PDF docs. We can search with regular expression from particular or whole pages. Aspose.PDF for Java

Extract Text from PDF using Java | Aspose.PDF for Java
https://docs.aspose.com/pdf/java/extract-text-from-pdf/
Extract the Text from PDF file is a common task for Java developers. Use the Aspose.PDF for Java Pdf library to extract text in just a few lines of code. Most PDF documents are not editable, making converting the PDF to text a tedious if not impossible task, especially if the solution involves bulk processing of PDF documents.

get text from pdf java - search.aspose.com
https://search.aspose.com/q/get-text-from-pdf-java-4.html
PDF Text Annotation | Aspose.PDF for Java,Get font style bold italic while extracting text from PDF using Aspose.PDF for .NET - Aspose.PDF Product Family - Search. Sort Score Result 10 results Languages All Labels All Results 31-40 of 8,167 for get text from pdf java (0.03 sec) ...

Extract Text from a PDF using Android Java - Knowledge ...
https://kbdeveloper.qoppa.com/extract-text-from-a-pdf-using-android-java/
Extract Text from a PDF using Android Java / Android PDF Toolkit - qPDF / Extract Text from a PDF using Android Java. January 10, 2017; Android PDF Toolkit - qPDF; Sample Android program to extract text content from a PDF document as a String using Qoppa's Android toolkit qPDF Toolkit. This program will extract the text from all pages of the PDF.



User #1's judgement



User #2's judgement



Beyond binary relevance

- With access to this kind of ranking scores, we will be able to roughly create the preferred rank list for each user
 - To build a personalized search engine
- Fundamental for the **learning to rank** approach
- Before that, we have to know more about the common evaluation metric for the case.
 - Discounted cumulative gain (DCG)

Discounted cumulative gain

- Multiple levels of relevance
- Two assumptions:
 - Highly relevant documents are more useful than marginally relevant documents.
 - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined.

Discounted cumulative gain

- Graded relevance is the measure of usefulness from examining a document
- Usefulness is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks
- Typical discount function is $1/\log_2(\text{rank})$
 - e.g., the discount at rank 4 and rank 8 are 1/2 and 1/4, respectively.

Discounted cumulative gain

- Let the relevance judgments be in a scale of [0, r] where r>2
- CG at rank n = $r_1 + r_2 + \dots + r_n$
- DCG at rank n = $r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \dots + r_n/\log_2 n$

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i}$$

Discounted cumulative gain — example

- 10 ranked document in 0-3 relevance scale:

3	2	3	0	0	1	2	2	3	0
---	---	---	---	---	---	---	---	---	---

- Discounted gain:

3	2/1	3/1.59	0	0	1/2.59	2/2.81	2/3	3/3.17	0
=	3	2	1.89	0	0	0.39	0.71	0.67	0.95

- DCG:

3	5	6.89	6.89	6.89	7.28	7.99	8.66	9.61	9.61
---	---	------	------	------	------	------	------	------	------

Normalized DCG

- Normalize DCG at rank n by the DCG value at rank n of the known-to-be best ranking
 - i.e., the ideal ranking — the ranking that perfectly order the documents by the degree of relevance
- Normalization is useful for demonstrating queries with varying numbers of relevant results

Normalized DCG

i	Ideal		Ranking function 1		Ranking function 2	
	Order	r _i	Order	r _i	Order	r _i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	NDCG _{ideal} = 1.00		NDCG _{function1} = 1.00		NDCG _{function2} = 0.9203	

$$DCG_{ideal} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{function_1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{function_2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{ideal} = 4.6309$$

It is all about telling whether users are happy

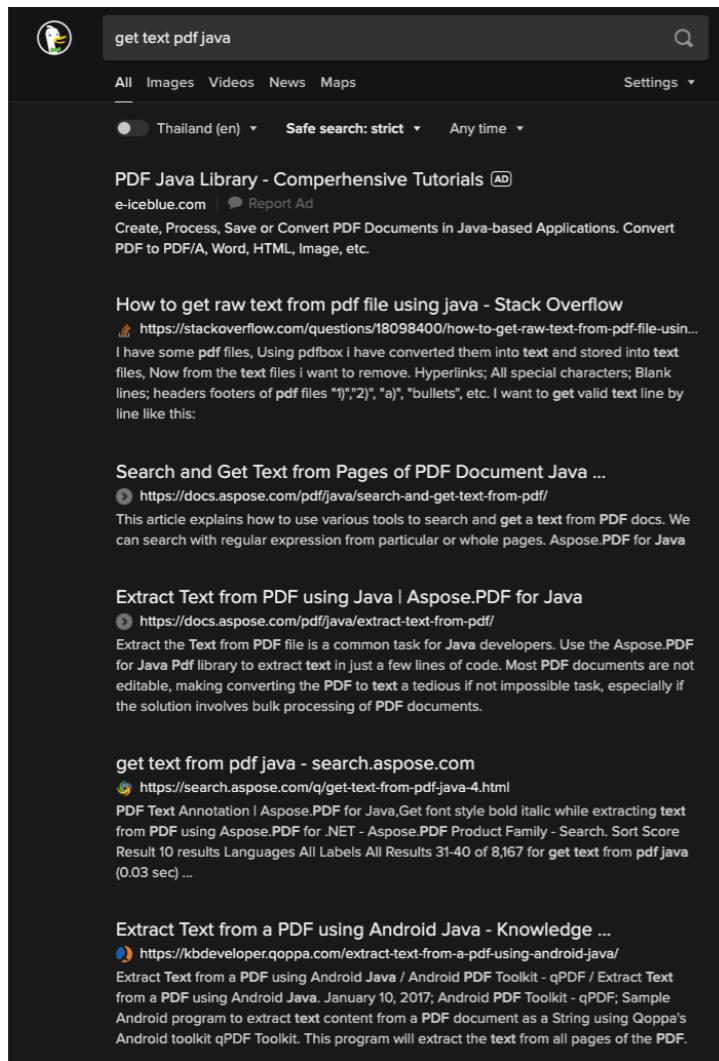
- Is the returned search products relevant to users?
- Do the users click the search results a lot?
- How much users spend a lot of money after searching for products using the search engine?
- How long does it take for a user to reach the solutions after starting the search

Why don't we measure user happiness directly

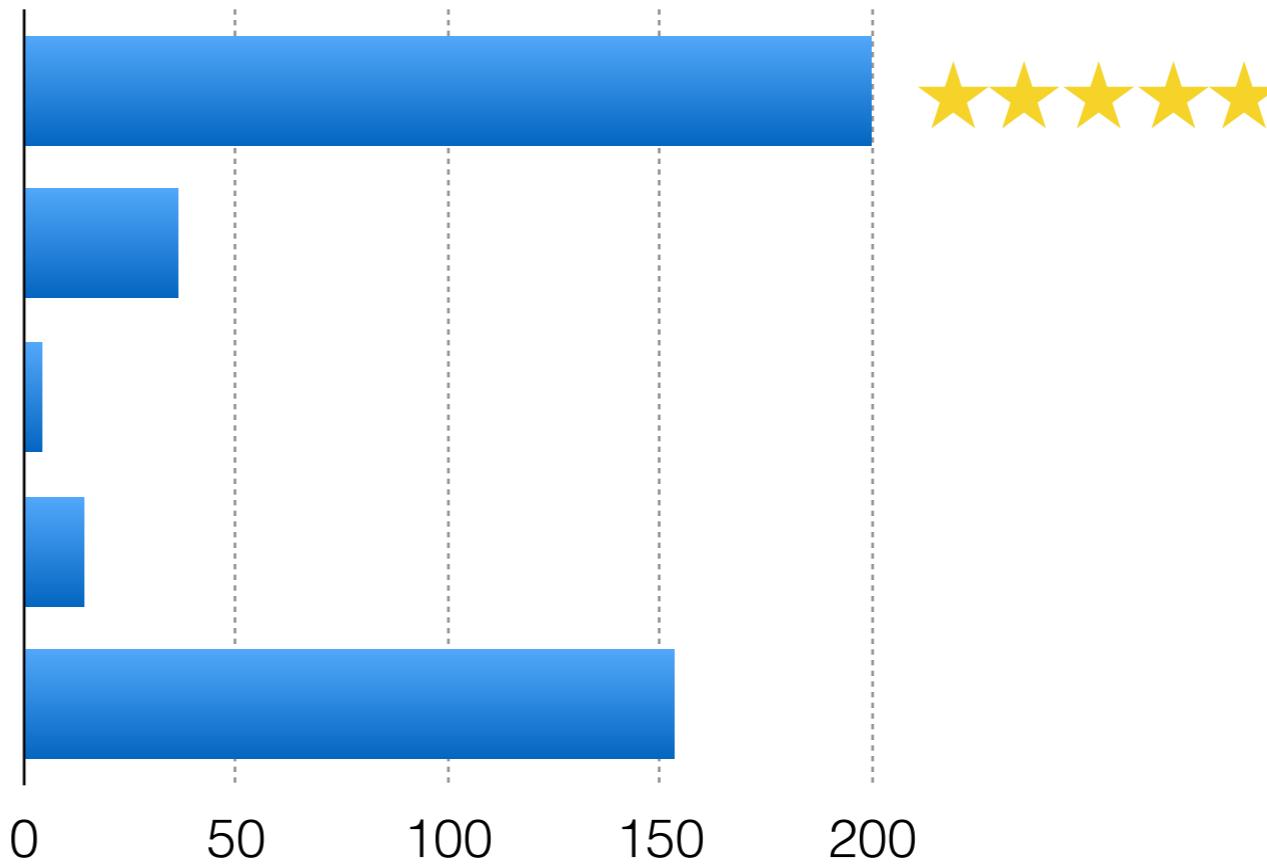
- Human judgments are
 - Subjective
 - Inconsistent — between raters and over time
 - Decay in value as documents | query evolves, i.e., how we answer lengthily questionnaire
 - Not easy to represent all real users
- Anyway, it is worth to give it a try

Measuring user clicks

- One of the most straightforward approaches to imply a user behavior learning

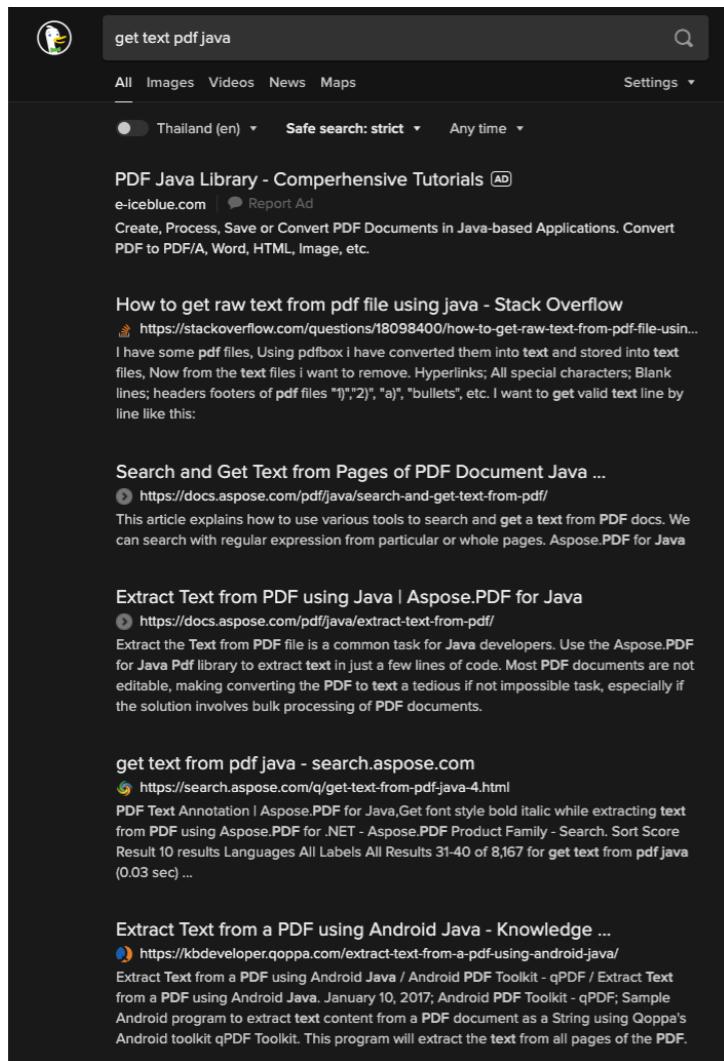


#clicks recorded

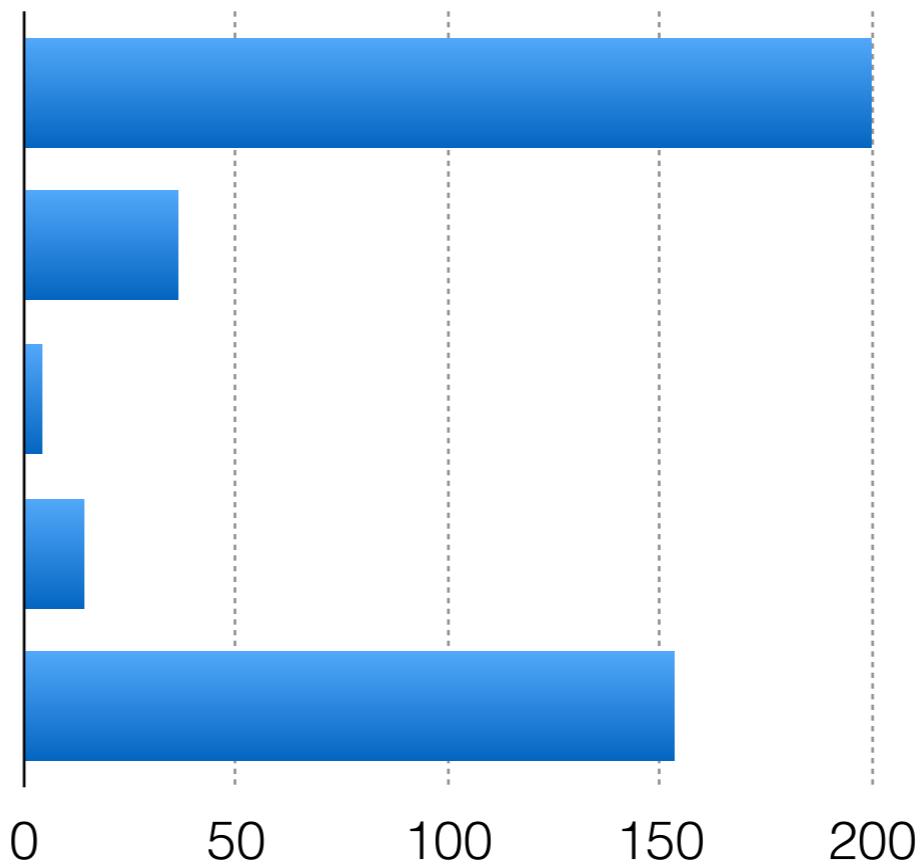


Exploiting the #user clicking information

- Adapt ranking to user clicks

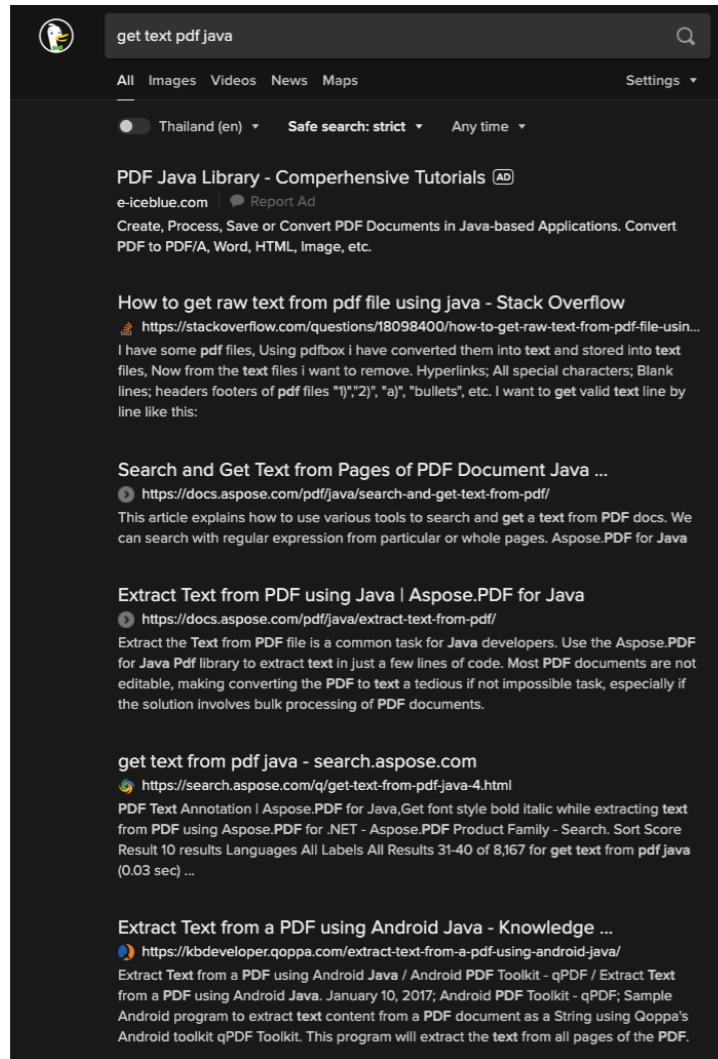


#clicks recorded

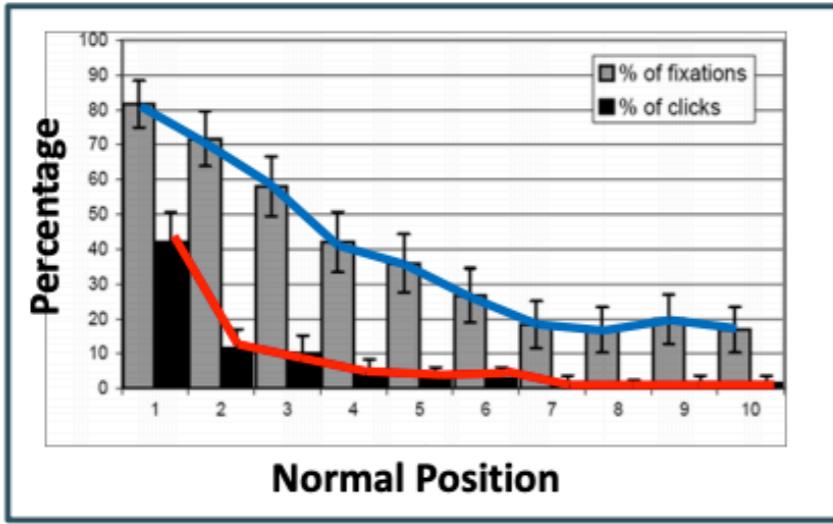


Eye-tracking

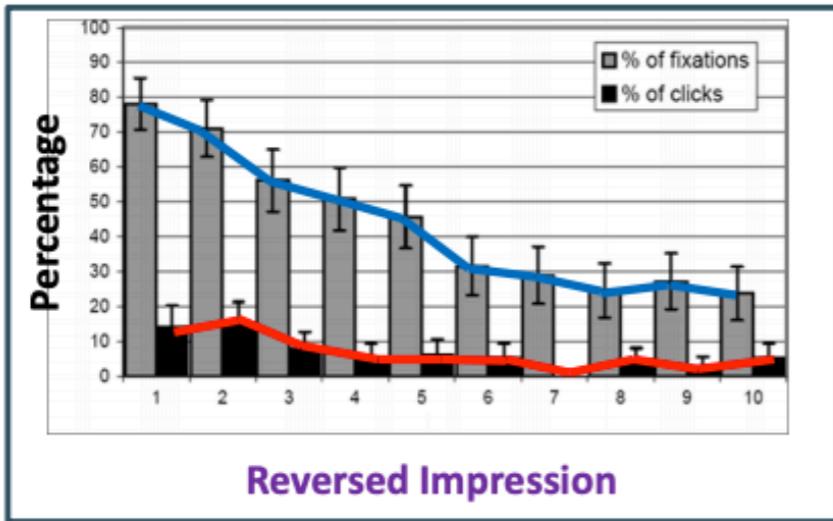
- One of the most straightforward approaches to imply a user behavior learning



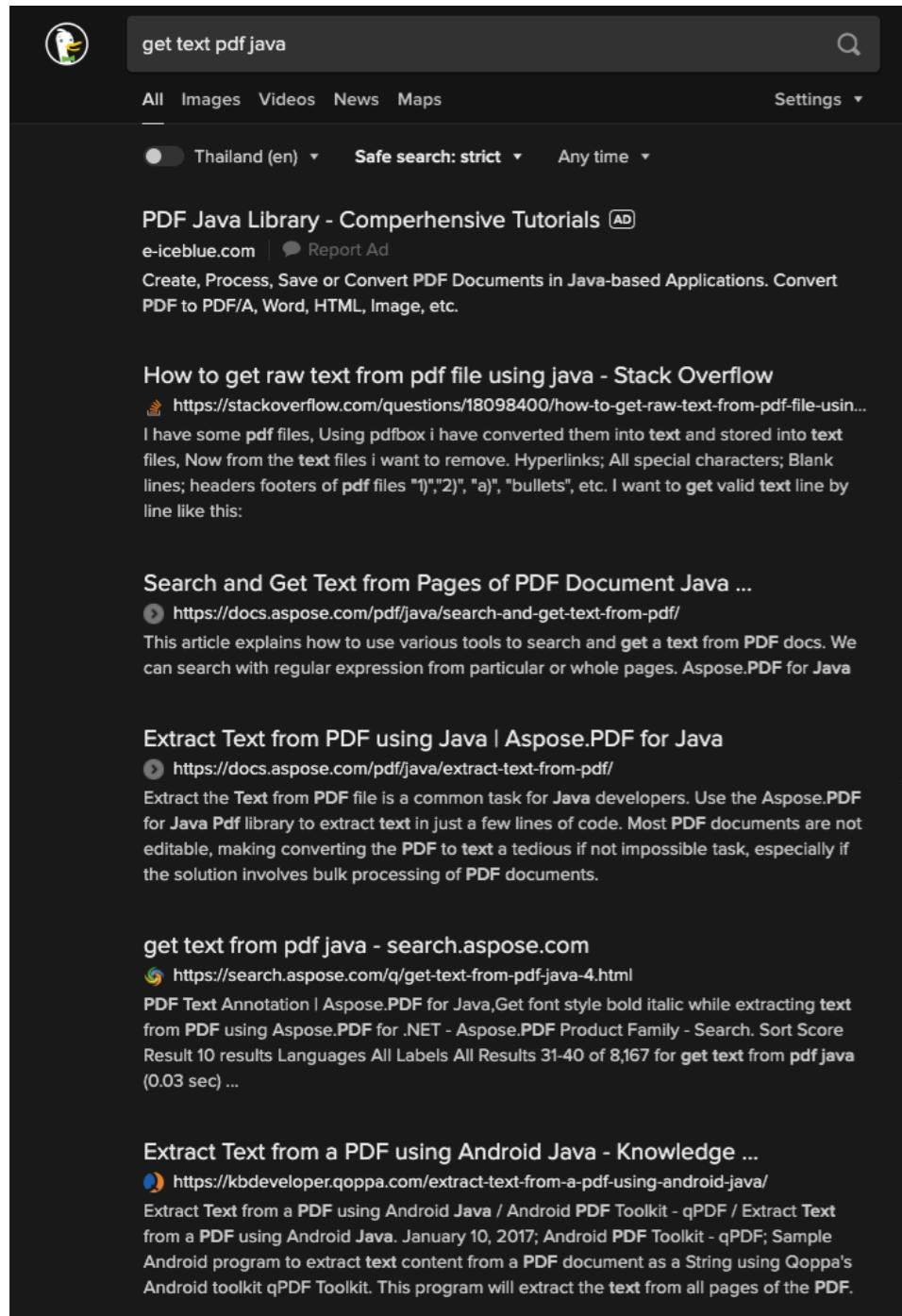
Informative but biased



- Higher positions receive more attention and more clicks
- Even the order is reversed



Relative rating



User's click sequence

- Hard to say Click #1 > Click #3
- Probably Click #3 > Click #2

Evaluating pairwise relative ratings

- Pairs of the form: Doc_a is better than Doc_b for a query q
- Not to assess a rank-ordering wrt per-doc relevance assessments but
 - to assess in terms of conformance with historical pairwise preferences recorded from user clicks

E.g., comparing two rankings via clicks

- Query: DevOps

Ranking #1

Wikipedia

Ultimate guide by AWS

CD&CI

Azure's service

Docker

Ranking #2

Docker

Jenkins

Intro to DevOps

CD&CI

Youtube's clip

E.g., comparing two rankings via clicks

- Interleaved ranking, this example start with #2



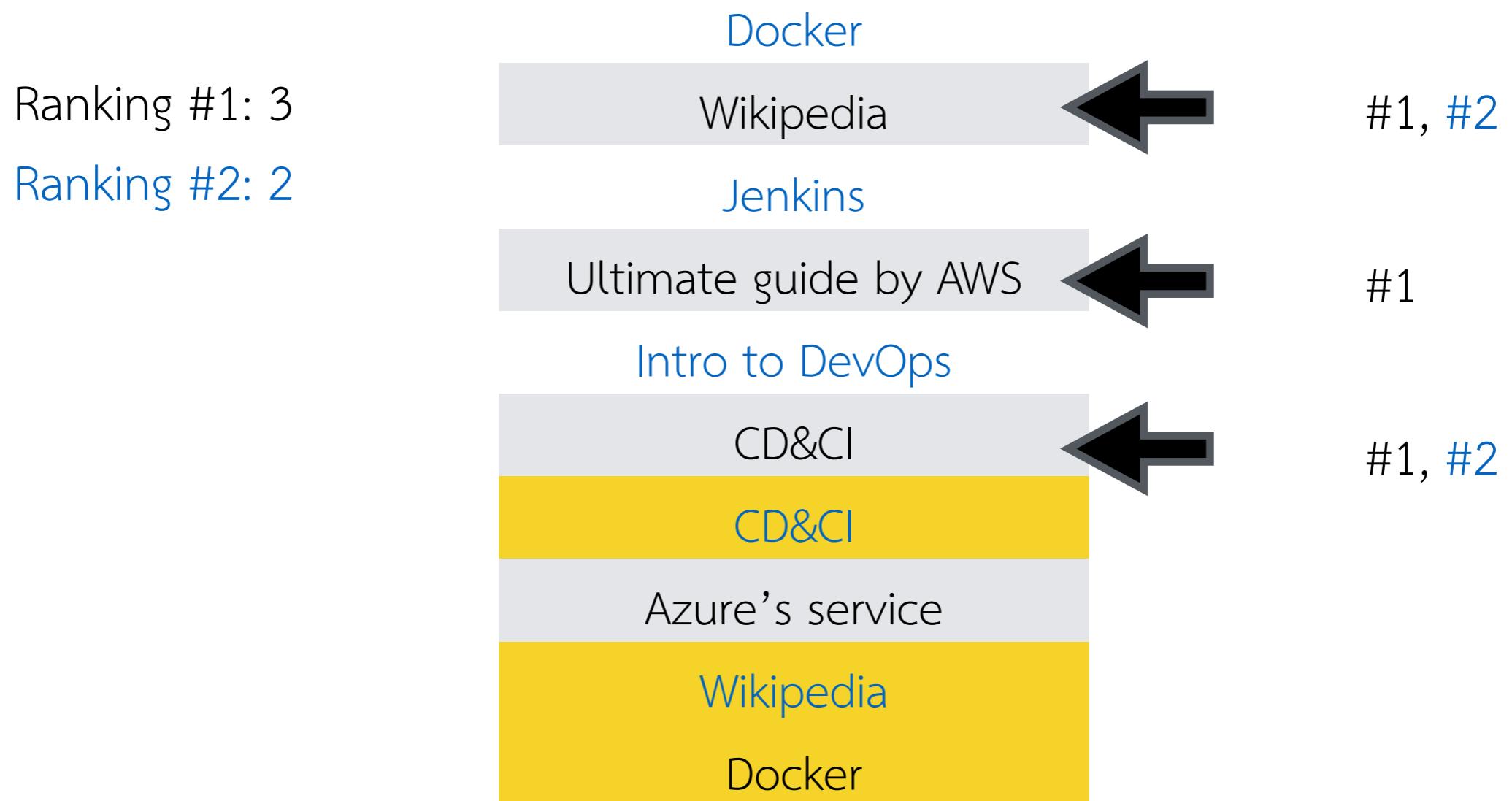
E.g., comparing two rankings via clicks

- Remove duplicate results



E.g., comparing two rankings via clicks

- Count user clicks



Interleaved ranking

- Present interleaved ranking to users
- Count clicks on results from #1 versus results from #2B
- Better ranking will (on average) get more clicks

A/B testing at web search engines

- To test new assumption about search results
- Prerequisite:
 - A large search engine is already up and running
 - Most users are using the old system
- What to do —
 - Divert a small portion of traffic to an experimental site and evaluate it.
 - Test can be done on either interleaved or full result.

What can be achieved from click profiling

devops

All Images Videos News Maps Settings ▾

Thailand (en) ▾ Safe search: moderate ▾ Any time ▾

DevOps Certification Training - Capstone Project in 3 Domains AD

simplilearn.com | Report Ad

Work on 20+ real-life projects on integrated labs & Capstone projects in 3 domains. Build Expertise in Configuration Management tools such as Puppet, SaltStack & Ansible

Courses: Project Management, Quality Management, Big Data & Analytics

What is DevOps? DevOps Explained | Microsoft Azure

<https://azure.microsoft.com/en-us/overview/what-is-devops/>

DevOps definition A compound of development (Dev) and operations (Ops), **DevOps** is the union of people, process, and technology to continually provide value to customers.

What does **DevOps** mean for teams?

What is DevOps? - Azure DevOps | Microsoft Docs

<https://docs.microsoft.com/en-us/devops/what-is-devops>

A compound of development (Dev) and operations (Ops), **DevOps** is the union of people, process, and technology to continually provide value to customers. What does **DevOps** mean for teams?

DevOps

DevOps is a set of practices that combines software development and IT operations. It aims to shorten the systems development life cycle and provide continuous delivery with high software quality. DevOps is complementary with Agile software development; several DevOps aspects came from the Agile methodology.

W More at Wikipedia

Share Feedback

User behavior

- User behavior is an intriguing source of relevance data:
 - Users make (somewhat) informed choices when they interact with search engines
 - Potentially a lot of data available in search logs
- Challenges
 - User behavior data can be very noisy
 - Interpreting user behavior can be tricky
 - Spam can be a significant problem
 - Not all queries will have user behavior

Time for questions