

# LC029 정보검색

2022 11 3

## Chapter 8 : Evaluation in information retrieval

# Measures for a Search Engine

# Measures for a Search Engine

가

- How fast does it index? ?
  - Number of documents/hour
- How fast does it search? 가 ?
  - Latency as a function of index size
- How large is its document collection? ( ) 가?
  - Many documents across a broad range of topics
- How expressive is its query language? 가 가?
  - Ability to express complex information needs (n )
  - Speed on complex queries
- How simple is the user interface? 가 가?

# Measures for a Search Engine

---

- All of the preceding criteria are ***measurable***.
  - We can quantify speed/size.
  - We can make expressiveness precise with feature checklists.
- The key measure is ***user's satisfaction***.
  - What is this?
  - Speed of response and size of index could be factors.
  - But blindingly fast, useless answers won't make a user satisfied.
- We need a way of quantifying user's satisfaction.

가가 가

?

# Measuring User's Satisfaction

---

## ■ Issues

- Who is the **user** we are trying to make satisfied?
- It depends on the setting:  
web engine users, eCommerce users, enterprise users

## 1. Web engine

- Users are satisfied when they find what they want.<sup>가</sup>
- If satisfied, they return to the same engine.
- Measuring user's satisfaction is measuring the rate of return users.<sup>가</sup>

# Measuring User's Satisfaction

---

## 2. eCommerce site

- eCommerce **site owners** vs. eCommerce **site users** ( ) 가 가
- We are trying to optimize eCommerce site owner's satisfaction because they pay us. 가
- Measuring user's satisfaction is measuring
  - Shoppers find what they want and make a purchase 가
  - Time to purchase
  - Fraction of searchers who become buyers : 가

# Measuring User's Satisfaction

---

## 3. Enterprise intranet search engine

- Such as company, government, academic : , ,
- They care about ***user productivity***:  
How much time do users save when looking for information?  
가?

# Measuring User's Satisfaction

---

- What is user's satisfaction?
  - Difficult to measure user's satisfaction.
  - So, standard methodology uses **relevance** of search results.
  - Then, how do we measure relevance?

가

가



# Relevance Judgment

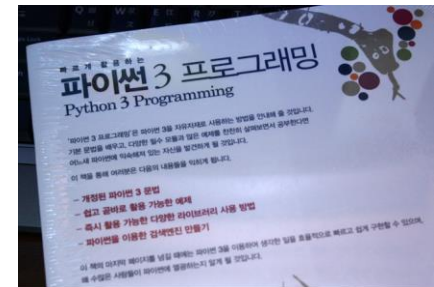
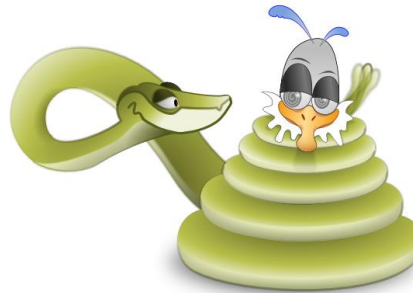
- Assessed relative to an **information need**, not a query

*Information on whether drinking red wine is more effective at reducing your risk of heart attacks than drinking white wine*

*wine AND red AND white AND heart AND attack AND effective*

- A document is relevant if it addresses the information need, not whether it has the query word.

**python**



# Test Collections for IR System Evaluation

가

# Information Retrieval System Evaluation

---

- Preparations:
  - A document collection : ( ) ,
  - A **test set** of information needs, expressible as queries
  - An assessment (usu. binary) of either ***Relevant*** or ***Nonrelevant*** for each query and each document
- Average performance over fairly large test sets (at least 50 information needs or more)
  - Why average?  
Because results are highly variable!

# Standard Test Collections

---

- Cranfield Collection

- 1398 abstracts of aerodynamics journal articles
- A set of 225 queries
  - Exhaustive relevance judgments of all (q, d) pairs.  
(by human experts)
- Pioneering test collection built in the late 1950s
  - Nowadays too small for the measure of IR effectiveness.
  - Thus, mostly used for elementary pilot experiments.

가

가  
( : )

# Standard Test Collections

---

- TREC (Text REtrieval Conference)

- 1.89 million documents (mostly newswire articles)
- Relevance judgments for 450 information needs.
  - No exhaustive relevance judgments.

- GOV2

- TREC test collection for use in the Terabyte Track.
- A large proportion of the crawlable pages in .GOV sites.
- 25 million web pages
  - Largest collection that is easily available.
  - But still 3 orders of magnitude smaller than what Google / Yahoo / MSN index.

# Standard Test Collections

---

- NTCIR (NII Test Collections for IR systems)<sup>TREC</sup>
  - Various test collections of similar sizes to TREC collection.
  - Focusing on East Asian Language and Cross-language IR.
- REUTERS-21578 and Reuters-RCV1
  - Reuters-21578 consists of 21,578 newswire articles.
    - Used for text classification.<sup>가 ( )</sup>
  - RCV1, much larger, consists of 800,000 documents.
- 20 Newsgroups<sup>== 가 가</sup>
  - Widely used text classification collection.
  - Consists of 1,000<sup>1000</sup> articles from 20 USENET newsgroups.
    - It contains 18,941 articles (after removing duplicates).

가

가

# Evaluation of **Unranked** Retrieval Sets

## Evaluation of **Ranked** Retrieval Results

가

가

# Evaluation of Unranked Retrieval Sets

---

- How about using **Accuracy** for evaluation? ?
  - An engine classifies each document as **relevant** or **nonrelevant** for a given query. /
  - The engine retrieves the documents classified as **relevant**.
- **Accuracy**
  - Fraction of correctly classified documents
  - A commonly used evaluation measure in machine learning classification work.
  - But, not a useful evaluation measure in IR.

Why?

: , 가 , 가 , 가 .



# Why not use accuracy in IR?

---

- How to build a 99.9% accurate search engine on a low budget....

99.9%가

The logo for Snoogle.com, featuring the word "snoogle" in a stylized, multi-colored font (blue, orange, and yellow) and ".com" in a smaller, blue font.

Search for:

*0 matching results found.*

# Why not use accuracy in IR?

---

- What we are searching for is a small portion of web docs.
- So, accuracy would be very high by classifying all documents as nonrelevant, if the 99.9% of the documents are in the nonrelevant category.
- However, people doing information retrieval *want to find something* and have a certain tolerance for junk.
- That is why we do not use **accuracy measure** in IR.
- Instead, we use **precision** and **recall** to evaluate the performance of IR systems.

# Precision and Recall

---

- **Precision**

- Fraction of retrieved docs that are relevant.

$$\frac{\#(\text{relevant docs retrieved})}{\#(\text{retrieved docs})} = P(\text{relevant} | \text{retrieved})$$

- **Recall**

- Fraction of relevant docs that are retrieved.

$$\frac{\#(\text{relevant docs retrieved})}{\#(\text{relevant docs})} = P(\text{retrieved} | \text{relevant})$$

# Precision and Recall

- **Precision** =  $\frac{\#(\text{relevant docs retrieved})}{\#(\text{retrieved docs})}$
- **Recall** =  $\frac{\#(\text{relevant docs retrieved})}{\#(\text{relevant docs})}$

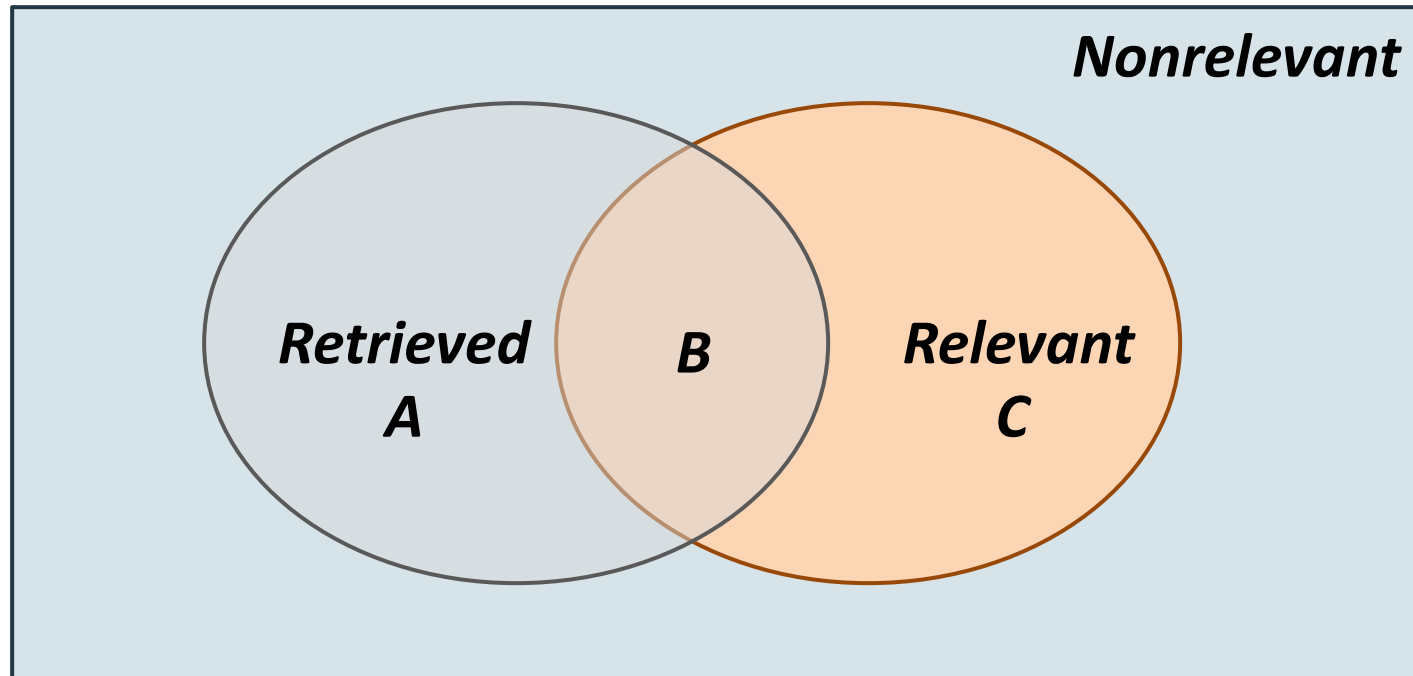
	Relevant	Nonrelevant
Retrieved	true positive( <b><i>tp</i></b> )	false positive( <b><i>fp</i></b> )
Not Retrieved	false negative( <b><i>fn</i></b> )	true negative( <b><i>tn</i></b> )

Precision  $P = tp / (tp + fp)$

Recall  $R = tp / (tp + fn)$

# Precision and Recall

---



$$\text{Precision} = B/A$$

$$\text{Recall} = B/C$$

# High Precision vs. High Recall

---

- Typical web surfers want every result on the first page to be relevant. : 가  
⇒ They want **high precision**.
- Some professional searchers (paralegals and intelligence analysts) want **high recall**. 가 : 가
- Individuals searching their Hard Disks want **high recall** (desktop search).

# Precision vs. Recall

---

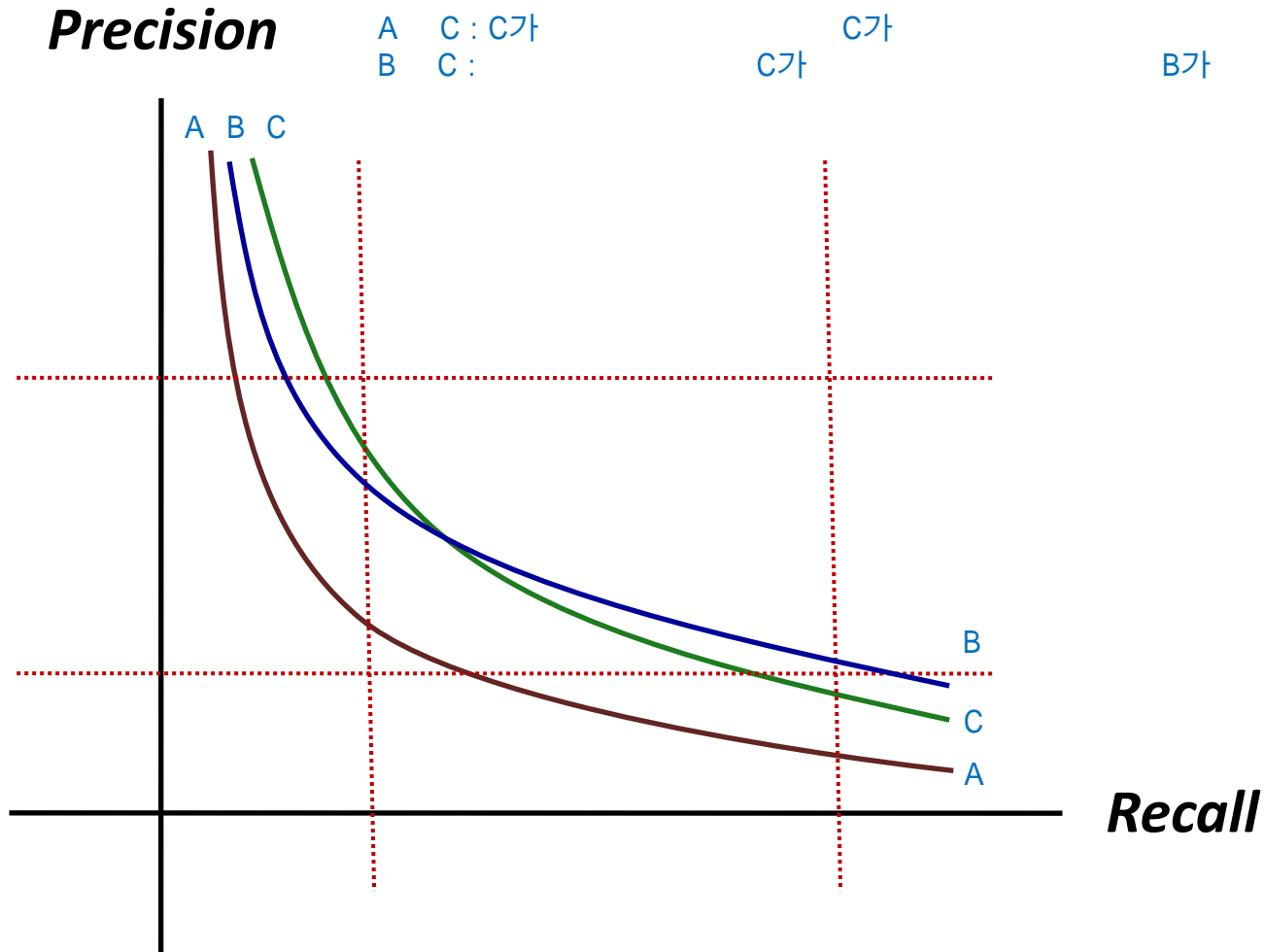
## ■ Precision

- Get high precision by retrieving just one relevant document .  
⇒ But the recall is very low.
- Precision decreases as the number of documents retrieved increases.

## ■ Recall

- Get high recall by retrieving all documents.  
⇒ But the precision is very low.
- Recall is a non-decreasing function of the number of documents retrieved.

# Precision-Recall Graph





# F measure : a combined measure

---

- F measure is a weighted **harmonic mean**:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{where } \beta^2 = \frac{1-\alpha}{\alpha}$$

- People usually use balanced  $F_1$  ( $F_{\beta=1}$ ) measure.  
When  $\alpha = \frac{1}{2}$  or  $\beta = 1$  (Note that  $0 \leq \alpha \leq 1$ )

$$F_{\beta=1} = \frac{2PR}{P + R}$$

If  $\alpha > \frac{1}{2}$  or  $\beta < 1$ , precision is emphasized.

If  $\alpha < \frac{1}{2}$  or  $\beta > 1$ , recall is emphasized.

# mean, median and mode

## ■ Mean

- Arithmetic mean :  $\frac{a+b}{2}$

- Geometric mean :  $\sqrt[n]{ab}$

- Harmonic mean :  $\frac{2ab}{a+b} = \frac{2}{\frac{1}{a} + \frac{1}{b}}$

$$\frac{a_1 + a_2 + \dots + a_n}{n}$$

$$\sqrt[n]{a_1 \cdot a_2 \cdot \dots \cdot a_n}$$

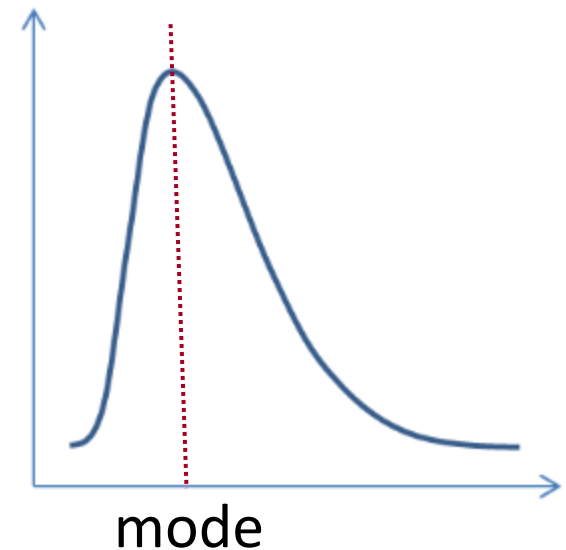
$$\frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_n}}$$

## ■ Median

- The central value of a series  
가

## ■ Mode

- The most frequent value in a series  
가



가 가

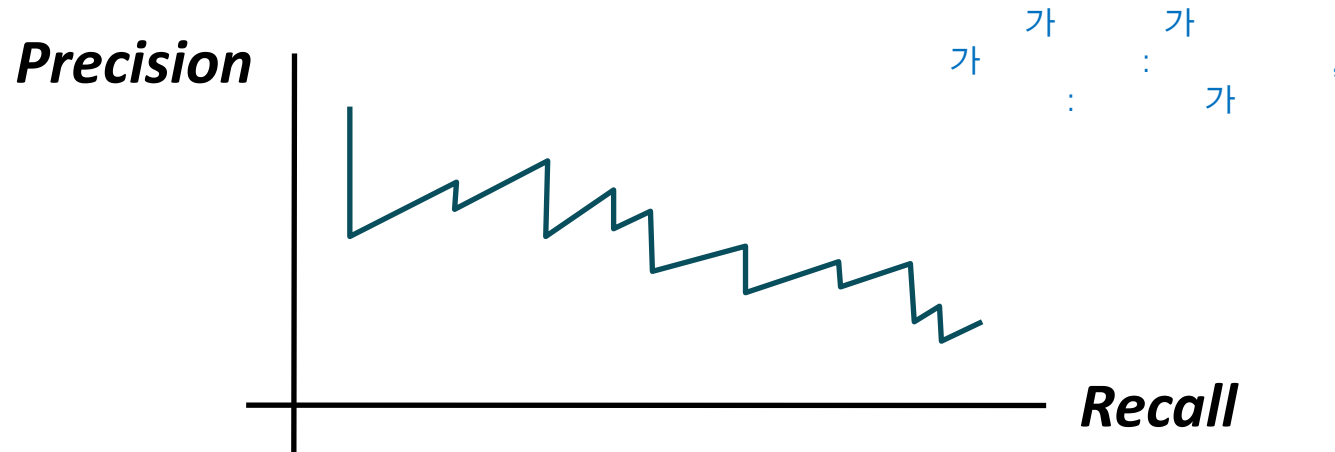
# Evaluation of Unranked Retrieval Sets

# Evaluation of **Ranked** Retrieval Results

가 가

# Evaluation of Ranked Retrieval Results

- Precision, Recall and F measure are set-based measures.
- In a ranked retrieval, the top  $k$  documents are presented.
  - If the  $(k+1)^{\text{th}}$  document retrieved is nonrelevant, the recall is the same as for the top  $k$  documents, while the precision has dropped.
  - If the  $(k+1)^{\text{th}}$  document retrieved is relevant, both precision and recall increase.



# Evaluation of Ranked Retrieval Results

---

가 5가

1. Interpolated Precision
2. 11-point Interpolated Average Precision
3. Mean Average Precision (MAP)
4.  $R$ -precision
5. Precision at  $k$

# 1. Interpolated Precision

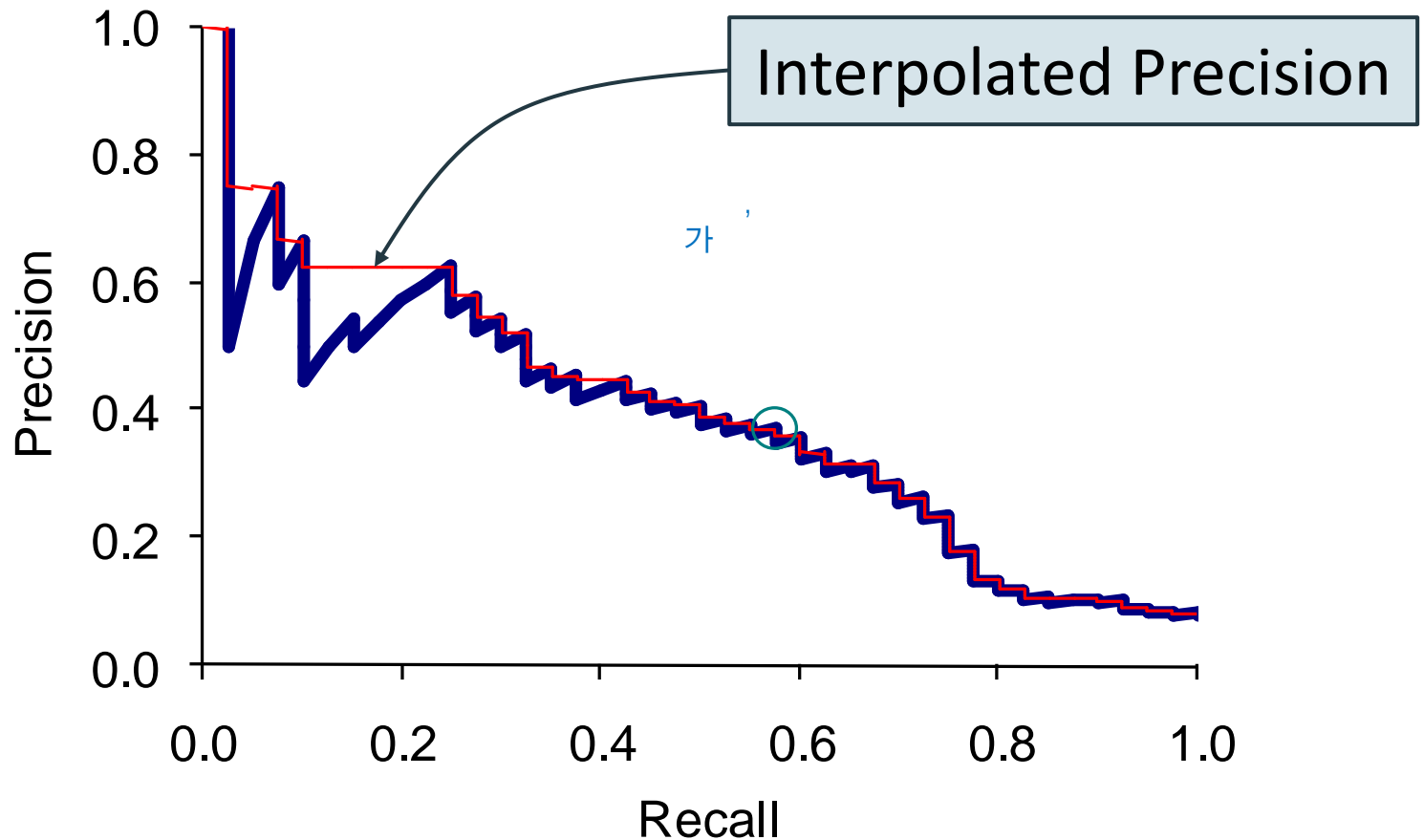
interpolate :

- **Interpolated precision** at a certain recall level  $r$  is defined as: recall

$$p_{interp}(r) = \max_{r' \geq r} p(r')$$

- Justification 가 ,
  - Almost anyone would be prepared to look at a few more documents if the precision of the larger set is higher.

# 1. Interpolated Precision



**A precision-recall curve**

## 2. 11-point Interpolated Average Precision

---

- Motivation
  - Graphs are good, but people want summary measures!
- 11-point interpolated average precision
  - The interpolated precisions are measured at the recall levels of 0.0, 0.1, ..., 0.9, 1.0.
  - At each recall level, the arithmetic mean of the **interpolated precision** is calculated for a set of queries.

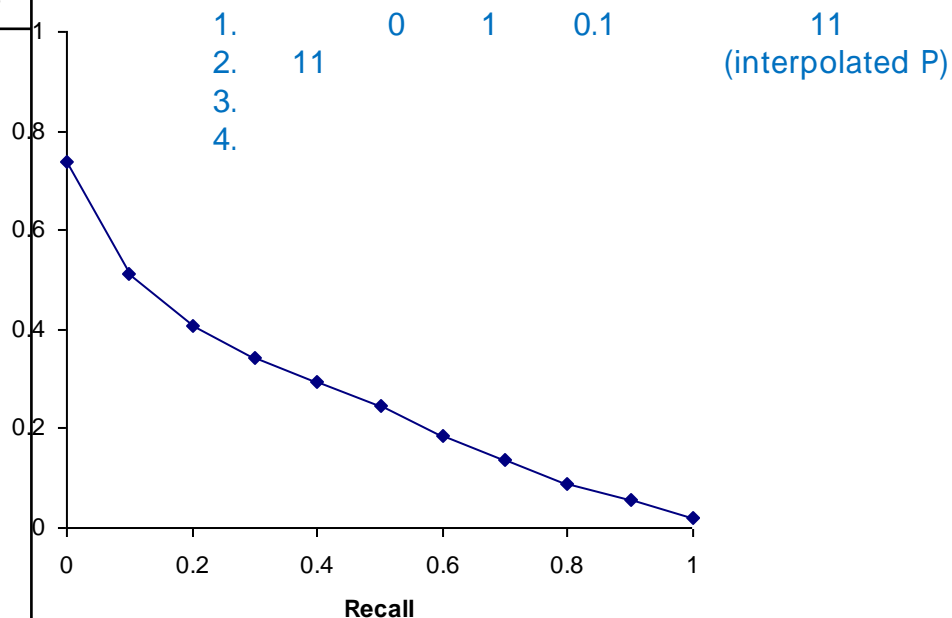
1.            0        1        0.1                    11  
2.    11    (interpolated P)  
3.  
4.

50            가            50    \* 11



## 2. 11-point Interpolated Average Precision

Recall	Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

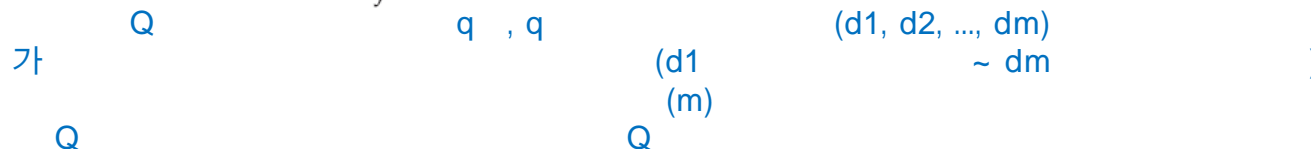


### 3. Mean Average Precision (MAP)

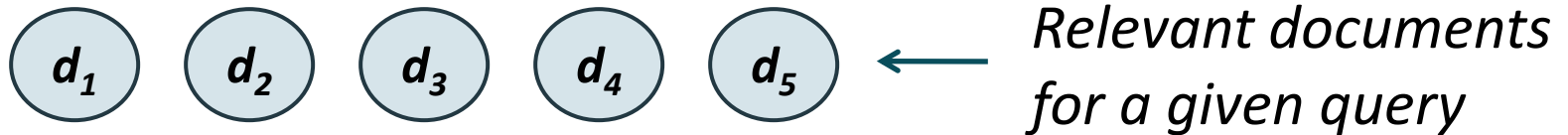
- Provide a **single-figure measure** of quality **across recall levels**.
- Let  $q_j \in Q$  be an information need and  $\{d_1, \dots, d_{m_j}\}$  be a set of relevant documents.
- $R_{jk}$  is the set of ranked retrieval results from the top result until you get to document  $d_k$  for  $q_j$

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

- When a relevant document is not retrieved at all, the  $\text{Precision}(R_{jk})$  in the above is taken to be 0.



### 3. Mean Average Precision (MAP)



$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

가 Q  
Q  
q, q  
Q  
(d1, d2, ..., dm)  
(m)  
~ dm  
)

### 3. Mean Average Precision (MAP)

---

- **MAP scores** normally **vary widely** across information needs when measured **within a single system**.
  - To measure system effectiveness, a set of test information needs must be large and diverse enough.
- **MAP is good to measure system effectiveness** for an individual information need **across systems**.

가

## 4. *R*-precision <sup>R</sup>

- Assume that we know in advance the number of relevant documents for a given query is *Rel*.<sup>가</sup>
- Of *Rel* documents retrieved, let *r* be the number of relevant documents.<sup>가</sup>  
$$R\text{-precision} = \frac{r}{Rel} \text{ (by definition, } R\text{-prececion} = \frac{r}{Rel} = \text{recall)}$$
- *R*-precision of a perfect system is 1.0.

10      가      가  
(      10 )  
, R

## 5. Precision at $k$

- Web surfers are interested in how many good results are included in the first page.
- So, we need to measure the precision of first  $k$  retrieved results.
  - Although the total number of relevant documents for a query has a strong influence on precision at  $k$ , the measure does not consider it.
    - R-precision considers the total number of relevant documents!!!
  - So, the measure is not stable and not commonly used in practice.

가

# Assessing Interjudge Agreement

## Kappa Measure

# Evaluation of large search engines

## A/B Testing

# Interjudge Agreement

---

- Human judgments are quite idiosyncratic and variable.
- So, we need a measure for inter-judge agreement.

가 가

가



# Interjudge Agreement: TREC 3

- Human agreement on a binary relevance judgment is quite modest.

query	# of docs judged	disagreement
51	211	6
62	400	157
67	400	68
95	400	110
127	400	106

- That is why we do not adopt more fine-grained relevance judgment.

- 가

# Interjudge Agreement

---

- So, build Test Collections
  - Search engines have test collections of queries and hand-ranked results.
  - Once we have test collections, we can **reuse** them (so long as we don't **overtrain** too badly).

( )

# Kappa Measure

가

- Agreement measure among 2 judges
- Designed for categorical judgments

$$\text{kappa} = \frac{P(A) - P(E)}{1 - P(E)} \quad \text{where}$$

$P(A)$  – proportion of judges agree 가

$P(E)$  – proportion of judges agree by chance 가

kappa = 1 for total agreement. 가

kappa = 0 for chance agreement. 가 가

kappa < 0 if they are worse than random. 가

# Kappa Measure: Example

# of docs	Judge 1	Judge 2
300	Relevant	Relevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant
70	Nonrelevant	Nonrelevant

$$P(A) = (300+70)/400 = 370/400 = \mathbf{0.925}$$

$$P_{judge1}(rel) = 320/400, P_{judge2}(rel) = 310/400 \quad \text{가 가 가}$$

$$P_{judge1}(\sim rel) = 80/400, P_{judge2}(\sim rel) = 90/400 \quad \text{가 가 가}$$

$$P(E) = P_{judge1}(rel) \cdot P_{judge2}(rel) + P_{judge1}(\sim rel) \cdot P_{judge2}(\sim rel) = \mathbf{0.665}$$

$$kappa = (0.925 - 0.665)/(1 - 0.665) = 0.776$$

가 0.776

# Kappa Measure

---

- $\text{kappa} > 0.8$   
good agreement
- $0.67 < \text{kappa} < 0.8$   
fair agreement (tentative conclusion)
- $\text{kappa} < 0.67$   
dubious assessment    가    ?
- The precise cutoff depends on purpose for which the data will be used
- For  $> 2$  judges    가    가    가  
average pairwise kappas

# Evaluation of large search engines

---

- Recall is difficult to measure on the web.
  - Search engines could use **precision at top  $k$** , e.g.  $k = 10$ .
  - Search engines also use **non-relevance-based measures**.
- Non-relevance-based measures
  - Clickthrough on first result
    - Not reliable if you look at a single clickthrough.
    - But pretty reliable in the aggregate.
  - A/B testing

# A/B Testing

---

- Used to test a *single* innovation of current system.
  - You have to keep a search engine running while testing the innovation of current system
- Let most users use old system.
- Divert a small proportion of traffic (say 1%) to the new system that includes the innovation.
- Evaluate with an automatic measure like *click stream mining* on first result.
  - To see if the innovation does improve user happiness.



# Result Snippets

# Result Summaries

---

- When presenting information retrieval results, most commonly, a list of the document titles plus a short summary is presented.

## [John McCain](#)

**John McCain 2008** - The Official Website of **John McCain's** 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...  
[www.johnmccain.com](http://www.johnmccain.com) · [Cached page](#)

## [JohnMcCain.com - McCain-Palin 2008](#)

**John McCain 2008** - The Official Website of **John McCain's** 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...  
[www.johnmccain.com/Informing/Issues](http://www.johnmccain.com/Informing/Issues) · [Cached page](#)

## [John McCain News- msnbc.com](#)

Complete political coverage of **John McCain**. ... Republican leaders said Saturday that they were worried that Sen. **John McCain** was heading for defeat unless he brought stability to ...  
[www.msnbc.msn.com/id/16438320](http://www.msnbc.msn.com/id/16438320) · [Cached page](#)

## [John McCain | Facebook](#)

Welcome to the official Facebook Page of **John McCain**. Get exclusive content and interact with **John McCain** right from Facebook. Join Facebook to create your own Page or to start ...  
[www.facebook.com/johnmccain](http://www.facebook.com/johnmccain) · [Cached page](#)

# Result Summaries

---

- The title is typically automatically extracted from document metadata. What about the summaries?
  - User can identify relevant documents based on title.
  - So, the title is crucial.

```
<head>  
<meta name="description" content="Free Web tutorials"/>  
<meta name="keywords" content="HTML,CSS,XML,JavaScript"/>  
<meta name="author" content="Hege Refsnes"/>  
<meta http-equiv="content-type" content="text/html; charset=UTF-8"/>  
</head>
```

# Result Summaries

---

- Two basic kinds of Summary:
  - A **static summary** of a document is always the same, regardless of the query that hit the document.
  - A **dynamic summary** is a *query-dependent* attempt to explain why the document was retrieved for the query at hand.

## [John McCain](#)

John McCain 2008 - The Official Website of John McCain's 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for McCain; Americans with ...  
[www.johnmccain.com](http://www.johnmccain.com) · [Cached page](#)

## [JohnMcCain.com - McCain-Palin 2008](#)

John McCain 2008 - The Official Website of John McCain's 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for McCain; Americans with ...  
[www.johnmccain.com/Informing/Issues](http://www.johnmccain.com/Informing/Issues) · [Cached page](#)

# 1. Static summaries

---

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic
  - Take the first 50 words of a document.
  - Particular zone of a document could be used.  
e.g. description zone of a document
  - Summary is cached at indexing time.

# 1. Static summaries

---

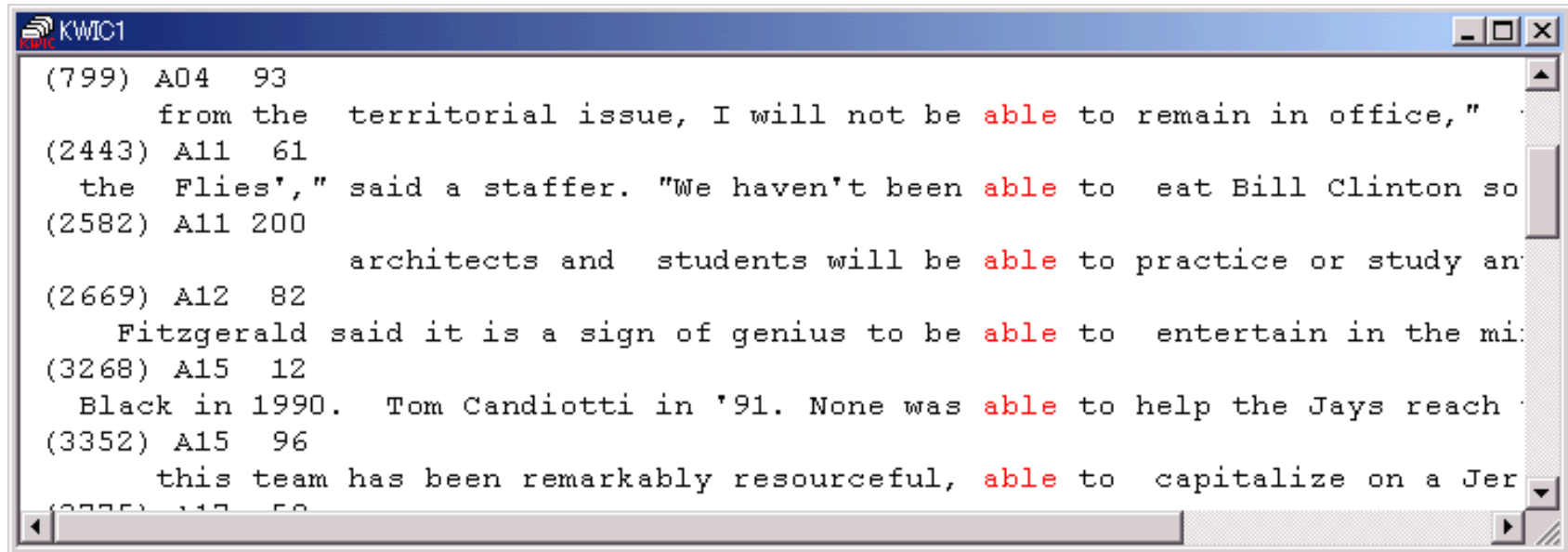
- More sophisticated
  - Extract from each document **a set of key sentences.**
  - Simple NLP heuristics to score each sentence.
    - Favoring first paragraph and last paragraph of a document.
    - Favoring first sentence and last sentence of a paragraph.
    - Favoring a sentence that include terms with high tf-idf scores.
    - Considering cue words (In conclusion, ..., For example, ...).
  - Summary is made up of top-scoring sentences.
- Most sophisticated
  - NLP technique is used to synthesize a summary.
  - Seldom used in IR

## 2. Dynamic summaries

- Present one or more “windows” within the document that contain some of the query terms.
  - Referred to as KWIC snippets : **K**eyword **i**n **C**ontext presentation

\$ kwic *able*

가



The screenshot shows a window titled "KWIC1" with a list of search results for the keyword "able". Each result line consists of a line number in parentheses, a document identifier (A04, A11, A12, A15), and a page number. The word "able" is highlighted in red in the text snippets. The results are as follows:

Line Number	Document	Page	Text Snippet
(799)	A04	93	from the territorial issue, I will not be able to remain in office,"
(2443)	A11	61	the Flies'," said a staffer. "We haven't been able to eat Bill Clinton so
(2582)	A11	200	architects and students will be able to practice or study an
(2669)	A12	82	Fitzgerald said it is a sign of genius to be able to entertain in the mi
(3268)	A15	12	Black in 1990. Tom Candiotti in '91. None was able to help the Jays reach
(3352)	A15	96	this team has been remarkably resourceful, able to capitalize on a Jer

## 2. Dynamic summaries

- Generated in conjunction with scoring
  - If query found as a phrase, all or some occurrences of the phrase in the document will be shown as summary.
  - If not, document windows that contain multiple query terms will be selected.



christppher manning

### Christopher Manning, Stanford NLP

Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University.

[nlp.stanford.edu/~manning/](http://nlp.stanford.edu/~manning/) - 12k - [Cached](#) - [Similar pages](#)

christopher manning machine translation

### Christopher Manning, Stanford NLP

Christopher Manning, Associate Professor of Computer Science and Linguistics, ... computational semantics, **machine translation**, grammar induction, ...

[nlp.stanford.edu/~manning/](http://nlp.stanford.edu/~manning/) - 12k - [Cached](#) - [Similar pages](#)

가(cache)

가



## 2. Generating dynamic summaries

---

- Dynamic summaries
  - greatly improve the usability of IR systems.
  - present a complication for IR system design.
    - But users really like snippets, even if they complicate IR system design.
  - cannot be pre-computed.
    - *Documents* are cached at index time
    - Most often, cache a fixed-size prefix of the document
    - Summaries can be constructed, cueing from hits found in the positional index.

## 2. Generating dynamic summaries

---

- Generating good dynamic summaries is a tricky optimization problem. 가
  - The screen space for presenting the summary is normally small and fixed.
  - We want short item, so as to show as many KWIC matches as possible, and perhaps other things like title.
  - We want snippets to be long enough to be useful.
  - We want linguistically well-formed snippets, since users prefer snippets that contain complete phrases.

# Alternative results presentations

- An active area of HCI research
  - Apple's Cover Flow for search results

