ครั้งที่ 9

การประเมินผลการค้นคืนข้อมูล

Why System Evaluation?

- There are many retrieval models, algorithms, systems, which one is the best?
- What is the best component for:
  - Ranking function (cosine, …)
  - Term selection (stopword removal, stemming…)
  - Term weighting (TF, TF-IDF,…)
- How far down the ranked list will a user need to look to find some/all relevant documents?

Difficulties in Evaluating IR Systems

- Effectiveness is related to the ***relevancy*** of retrieved items
- Even if relevancy is binary, it can be a difficult judgment to make
- Relevancy, from a human standpoint, is:
  - Subjective: Depends upon a specific user's judgment
  - Situational: Relates to user's current needs
  - Cognitive: Depends on human perception and behavior
  - Dynamic: Changes over time

Human Labeled Corpora

- Start with a corpus of documents
- Collect a set of queries for this corpus
- Have one or more human experts exhaustively label the relevant documents for each query
- Typically assumes binary relevance judgments
- Requires considerable human effort for large document/query corpora

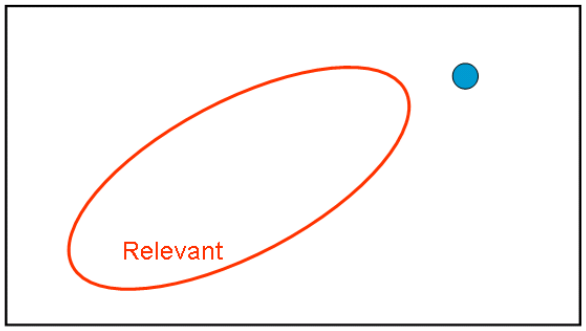
ประเด็นการประเมินผล

- ประสิทธิผลของการค้นคืนข้อมูล (Retrieval Effectiveness)
- ความสามารถของระบบในการที่จะให้บริการสารสนเทศตามที่ผู้ใช้ต้องการ
- วัดประสิทธิผลด้วยค่า Recall และ Precision

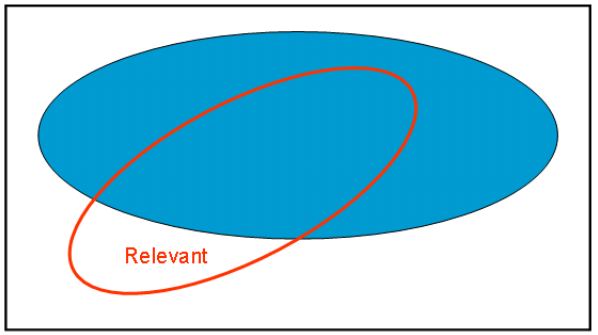
- ประสิทธิภาพ (Efficiency)
- วัดได้จากทรัพยากรของระบบคอมพิวเตอร์ที่ใช้ เช่น เนื้อที่เก็บความจำและ CPU Time เป็นต้น
- วัดได้จากค่าใช้จ่ายในการปฏิบัติการของระบบ, ความพยายามของผู้ใช้ (User Effort) ในการที่จะทำให้การค้นการเป็นไปอย่างมีประสิทธิภาพ

ประสิทธิผลของการค้นคืนข้อมูล

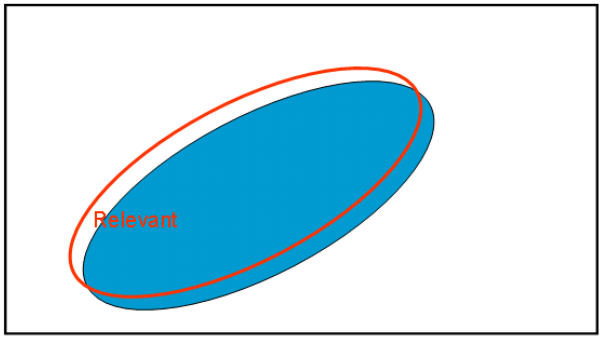
Very low precision, very low recall (0 in fact)

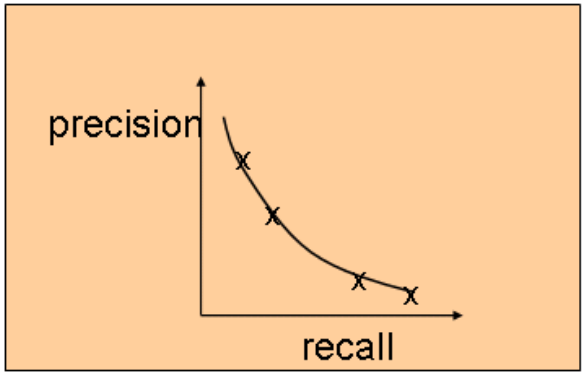


High recall, but low precision



High precision, high recall (at last!)





สาเหตุของการประเมินผล

- ต้องการที่จะทราบว่า เมื่อส่วนประกอบบางส่วนของระบบได้เปลี่ยนไป System Performance จะเปลี่ยนแปลงไปมากน้อยแค่ไหน

- ต้องการที่จะเปรียบเทียบความสามารถของระบบที่มีอยู่กับระบบค้นคืนสารสนเทศอื่นๆ

- ต้องการจำลอง (Simulate) การปฏิบัติการของระบบใหม่ก่อนที่จะมีการสร้างระบบจริงขึ้นมา

ส่วนประกอบของการทดสอบระบบ

- รายละเอียดเกี่ยวกับระบบและส่วนประกอบต่างๆ ของระบบหรือตัวแบบของระบบที่จะถูกตรวจสอบ
- ข้อสมมุติฐาน (Hypothesis) ชุดหนึ่งที่จะถูกทดสอบ หรือต้นแบบ (Prototype) เกี่ยวกับตัวแบบที่จะถูกวัด
- หลักเกณฑ์ที่จะแสดงถึงจุดประสงค์ของการทำงานของระบบและวิธีการที่จะวัดเกณฑ์ของการทำงานของระบบเป็นปริมาณตัวเลข
- วิธีการที่จะได้มาซึ่งข้อมูลและวิธีการประเมินผล

การทดสอบระบบ

- หลักเกณฑ์การวัดที่เรียกว่า Objective Quantitative Criteria
- ซึ่งสามารถวัดเป็นปริมาณตัวเลขเห็นได้ชัด

- ต้องการตรวจดู Dialog System
- กำหนดว่า ความเร็วในการค้นหาเวลาที่ยาวนานที่สุดในการที่จะส่งคำตอบกลับมา (Maximum Allowable Response Time) = 45 วินาที
- วิธีการประเมินผล ทำได้โดยถามผู้ใช้ฐานข้อมูลแต่ละคน เพื่อวัด Response Time ด้วยนาฬิกาจับเวลา

รวบรวมข้อมูลโดย

- การบันทึก
- การสังเกตโดยตรง
- แบบสอบถาม
- เทคนิคในการสัมภาษณ์
- ใช้วิธีการทางเครื่องกลในการรวบรวมข้อมูลที่ต้องการ

- จะต้องมีการเลือกพารามิเตอร์ ที่สำคัญต่อวัตถุประสงค์ของการประเมินผล
- สามารถถูกวัดได้โดยง่าย
- และมีความสัมพันธ์เกี่ยวเนื่องกัน

พารามิเตอร์

- ขนาดของกลุ่มเอกสาร
- System Response Time
- ความสามารถของระบบที่จะดึงเอกสารที่เกี่ยวข้อง
- การแสดงรูปแบบ (Representation) ของเอกสาร
- วิธีการค้นหา
- คุณสมบัติของผู้ใช้
- จำนวนเฉลี่ยของเอกสารที่เกี่ยวข้องกับข้อคำถาม
- จำนวนเฉลี่ยของเอกสารที่เกี่ยวข้องที่ถูกดึงโดยระบบ

Subjective Relevance Measure

- *Novelty Ratio*: The proportion of items retrieved and judged relevant by the user and of which they were previously unaware
  - Ability to find ***new*** information on a topic
- *Coverage Ratio*: The proportion of relevant items retrieved out of the total relevant documents ***known*** to a user prior to the search
  - Relevant when the user wants to locate documents which they have seen before (e.g., the budget report for Year 2000)

Other Factors to Consider

- *User effort*: Work required from the user in formulating queries, conducting the search, and screening the output
- *Response time*: Time interval between receipt of a user query and the presentation of system responses
- *Form of presentation*: Influence of search output format on the user's ability to utilize the retrieved materials
- *Collection coverage*: Extent to which any/all relevant items are included in the document corpus

ส่วนประกอบของระบบที่เกี่ยวข้องกับการประเมินผล

พารามิเตอร์ที่เกี่ยวข้องกับนโยบายของอินพุท

- อัตราความผิดพลาดและการเสียเวลา (Time Delays) ในการที่จะใส่เอกสารใหม่ในกลุ่มเอกสาร
- ความล่าช้า(Time Lag) ตั้งแต่การรับเอกสารหนึ่งที่กำหนดให้มาจนกว่าเอกสารนั้นจะปรากฏในแฟ้มข้อมูล
- Collection Coverage ซึ่งเป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ถูกบรรจุจริง ๆ ในแฟ้มข้อมูล

รูปแบบทางกายภาพของอินพุท (Physical Input Form)

- ความหมายของเอกสาร
- รูปแบบของเอกสารนั้น อาจจะอยู่ในรูปแบบ
  - ชื่อเรื่อง (Title)
  - บทย่อ (Abstract)
  - บทสรุป (Summary) ,
  - ข้อความเต็ม (Full Text)

ซึ่งจะมีผลต่อการสร้างดัชนีและการค้นหาเอกสาร นอกจากนี้ยังมีผลต่อค่าใช้จ่ายของระบบด้วย

ส่วนประกอบของระบบที่เกี่ยวข้องกับการประเมินผล

โครงสร้างของแฟ้มข้อมูลที่ใช้ในการค้นหา

(Organization of Search Files)

- มีผลต่อขั้นตอนของการค้นหาข้อมูล
- Response Time
- ความพยายามของผู้ควบคุมระบบ
- ประสิทธิภาพของระบบดึงข้อมูล

ภาษาดัชนี (Indexing Language)

- ภาษาดัชนีประกอบด้วยเซตของดัชนีและกฎต่าง ๆ ที่ถูกใช้เพื่อกำหนดดัชนีเหล่านี้ ให้กับเอกสารและข้อคำถามในการค้นหาข้อมูล
- ในระหว่างขั้นตอนของการสร้างดัชนี ค่าที่เหมาะสมกับการแสดงเนื้อหาของเอกสารจะถูกเลือกจากภาษาดัชนีและกำหนดให้กับรายการเอกสารตามกฎที่กำหนดขึ้นมา
- พารามิเตอร์ที่สำคัญเกี่ยวกับเรื่องนี้ก็คือ Exhaustivity และ Specificity ของภาษาดัชนี

ส่วนประกอบของระบบที่เกี่ยวข้องกับการประเมินผล

วิธีการสร้างดัชนี (Indexing Operation)

- ถ้าการสร้างดัชนีถูกกระทำด้วยมือคนที่ได้รับการอบรมมาอย่างดี ตัวแปรที่จะมีผลกระทบต่อวิธีการสร้างดัชนีนั้น ยังมีความสัมพันธ์กับอิทธิพลของประสบการณ์ของผู้สร้างดัชนีต่อ Performance และความถูกต้องของ Assigner Terms

การวิเคราะห์คำถาม (Question Analysis)

- เป็นการยากในการกำหนดดัชนีให้กับ Information Requests การสร้าง Boolean Statements ที่มีความหมาย และการเปรียบเทียบข้อเรียกร้องที่ถูกวิเคราะห์แล้วกับข่าวสารที่ถูกเก็บบันทึกนั้น ทั้งหมด

ส่วนประกอบของระบบที่เกี่ยวข้องกับการประเมินผล

วิธีการค้นหา (Search Operations)

- ชนิดของโครงสร้างแฟ้มข้อมูลที่ถูกนำมาใช้
- การเปรียบเทียบข้อคำถามกับเอกสาร
- อิทธิพลของกลวิธีการค้นหาข้อมูลต่อ Response Timeและบน Search Performance
- มาตรฐานความเกี่ยวข้องผู้ใช้ระบบ (หมายถึง มาตรฐานที่ผู้ใช้ระบบใช้ในการพิจารณาว่าผลลัพธ์ที่ได้จากการค้นหานั้นเกี่ยวข้องกับเรื่องที่ผู้ใช้ต้องการมากน้อยแค่ไหน)

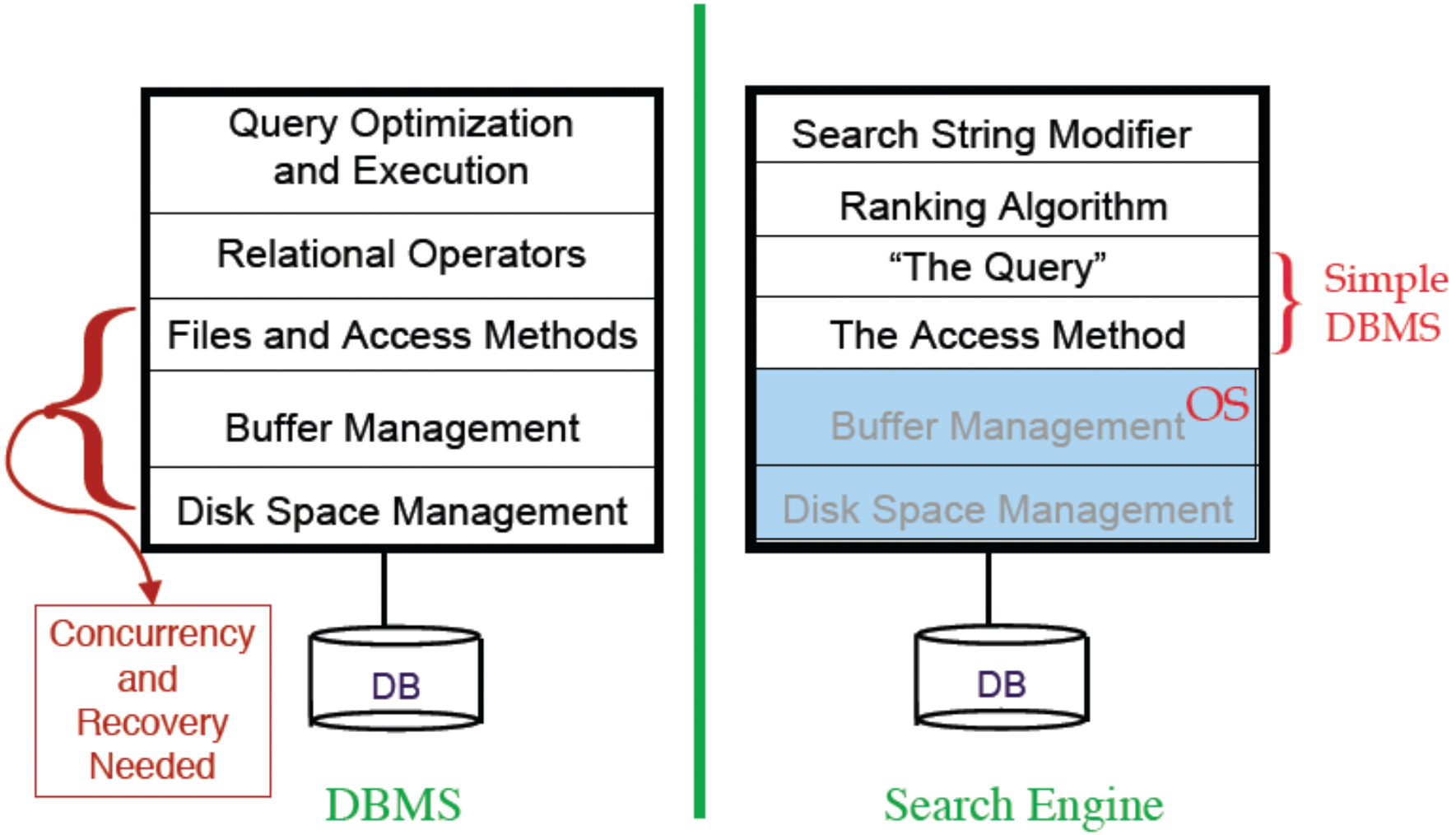
รูปแบบของการแสดงผลลัพธ์

- รูปแบบผลลัพธ์ที่ปรากฏออกมาจะมีผลต่อปริมาณความพยายามของผู้ใช้ที่จะดูผลลัพธ์ที่ได้จากค้นหาข้อมูลและความพอใจครั้งสุดท้ายที่ได้รับจากการค้นหาข้อมูลในแต่ละครั้ง

หลักเกณฑ์การประเมินผลเกี่ยวกับผู้ใช้

- ระบบควรที่จะสนองต่อความต้องการของผู้ใช้
- ข้อผิดพลาดในการดึงเอกสารที่เกี่ยวข้อง
- ชนิดของผลลัพธ์ที่ได้จากการค้นหาข้อมูล (เช่น เป็นชื่อเรื่อง, บทย่อ หรือข้อความเต็ม)เป็นต้น

Recall From the First Lecture



Collection Coverage

- เป็นปริมาณที่ทุก ๆ เอกสารที่เกี่ยวข้องถูกครอบคลุมอยู่ในระบบ
- หรืออาจจะกล่าวได้ว่า เป็นอัตราส่วนของจำนวนเอกสารที่เกี่ยวข้อง ต่อจำนวนเอกสารทั้งหมดในระบบ
- พารามิเตอร์ที่เกี่ยวข้อง ได้แก่
  - ชนิดของอุปกรณ์นำข้อมูลเข้า
  - ชนิดและขนาดของอุปกรณ์เก็บความจำ
  - ความยากง่ายของการวิเคราะห์เนื้อหา

การคำนวณ Recall และ Precision

Recall

- ถูกกำหนดให้เป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ถูกดึงออกมาจากจำนวนเอกสารที่เกี่ยวข้องทั้งหมด

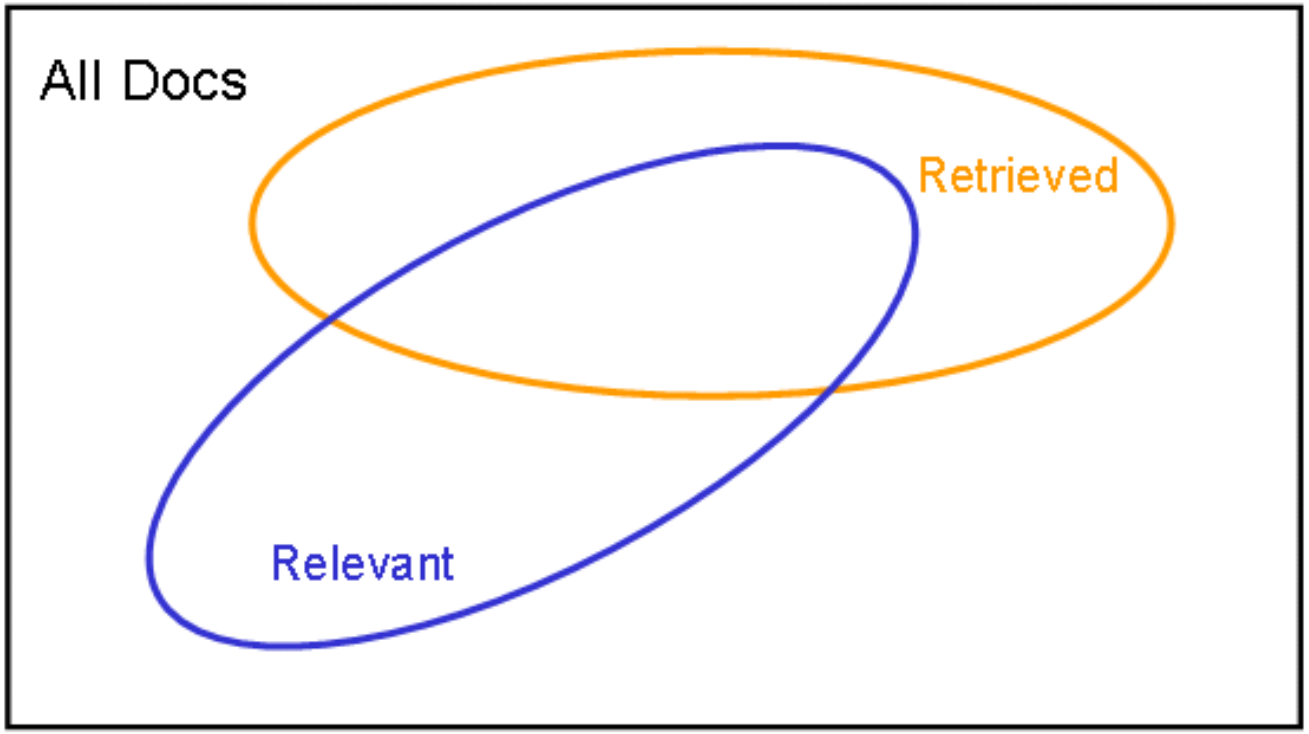
Precision

- เป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ถูกดึงออกมาจากจำนวนเอกสารที่ถูกดึงออกมาทั้งหมด

Partition of Collection

$$\text{Precision} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Retrieved}\,|}$$

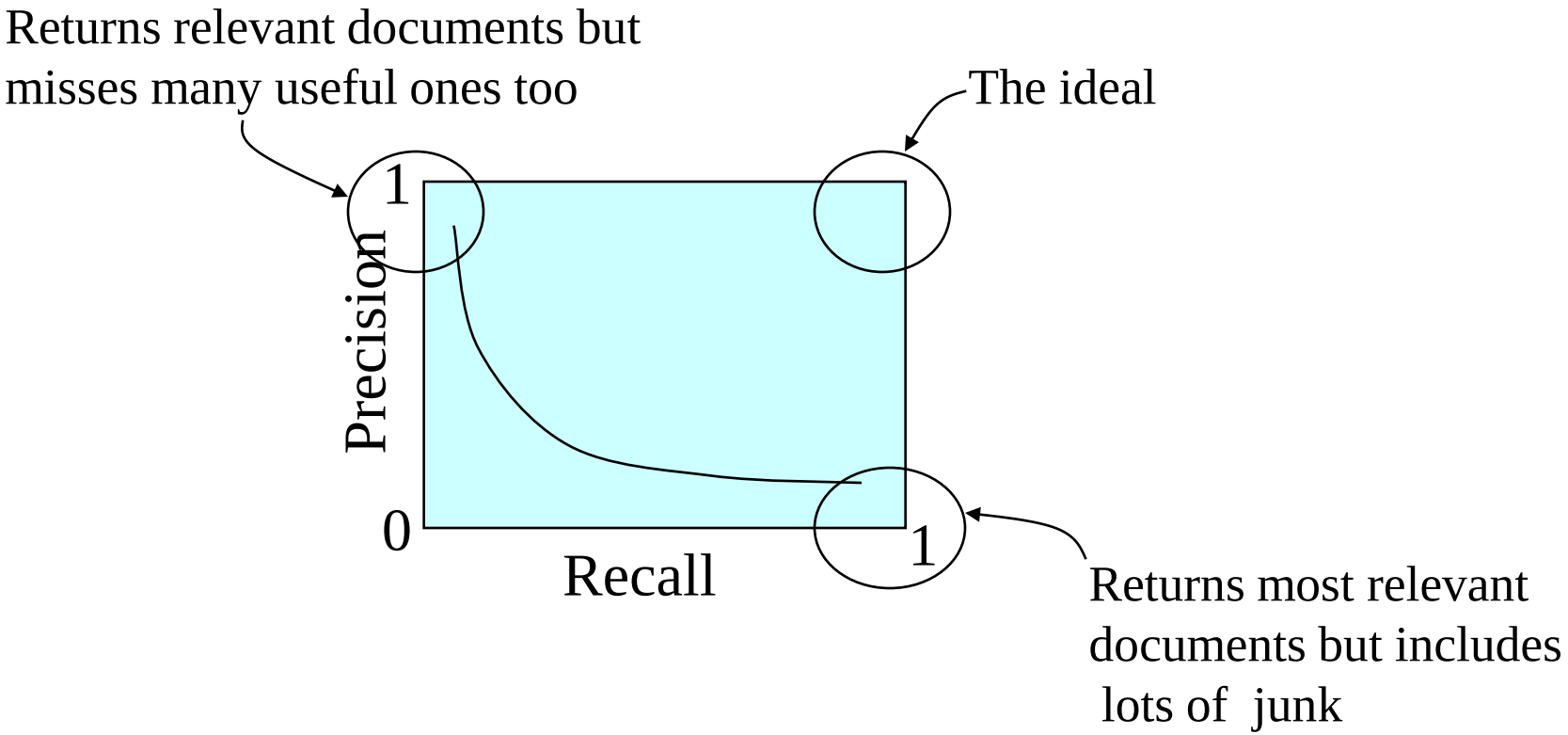
$$\text{Recall} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Rel in Collection}\,|}$$



Determining Recall is Difficult

- Total number of relevant items is sometimes not available:
  - Sample across the database and perform relevance judgment on these items
  - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set

Trade-off between Recall and Precision



Recall และ Precision

- สารสนเทศที่ต้องการของผู้ใช้แต่ละคนนั้นย่อมแตกต่างกันไป ซึ่งผู้ใช้บางคนอาจจะต้องการ Recall ที่สูง, กล่าวคือ, ทุกสิ่งที่ถูกดึงออกมาเป็นเรื่องที่น่าสนใจ
- ในขณะที่อีกคนหนึ่งอาจจะต้องการ Precision ที่สูง, กล่าวคือ, เอกสารที่ถูกดึงออกมามีเอกสารที่ไม่เกี่ยวข้องน้อยที่สุด
- ระบบค้นคืนสารสนเทศที่ดี ควรเป็นระบบที่มีทั้ง Recall และ Precision ที่สูง

Computing Recall/Precision Points

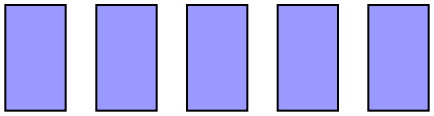
- For a given query, produce the ranked list of retrievals
- Adjusting a threshold on this ranked list produces different sets of retrieved documents, and therefore different recall/precision measures
- Mark each document in the ranked list that is relevant
- Compute a recall/precision pair for each position in the ranked list that contains a relevant document

Common Representation

	Relevant	Not Relevant
Retrieved	**A**	**B**
Not Retrieved	**C**	**D**

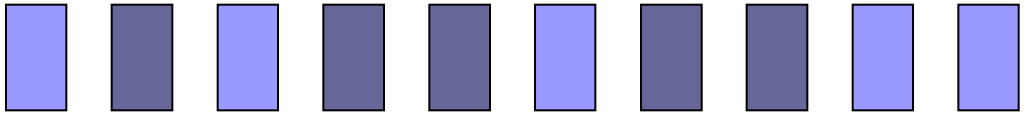
- Relevant = A+C
- Retrieved = A+B
- Collections size = A+B+C+D
- Precision = A/(A+B)
- Recall = A/(A+C)
- Miss = C/(A+C)
- False alarm = B/(B+D)

Precision and Recall Example



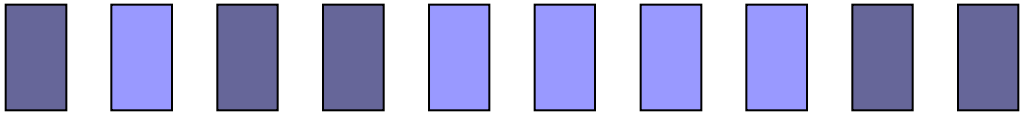
<- Relevant documents

Ranking 1



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

Ranking 2



Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

Average Precision of a Query

- Often want a single number effectiveness measure
  - E.g. for machine learning algorithm to detect improvement
- Average precision is widely use in IR
- Calculate by averaging when recall increases

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 62.2 %

Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

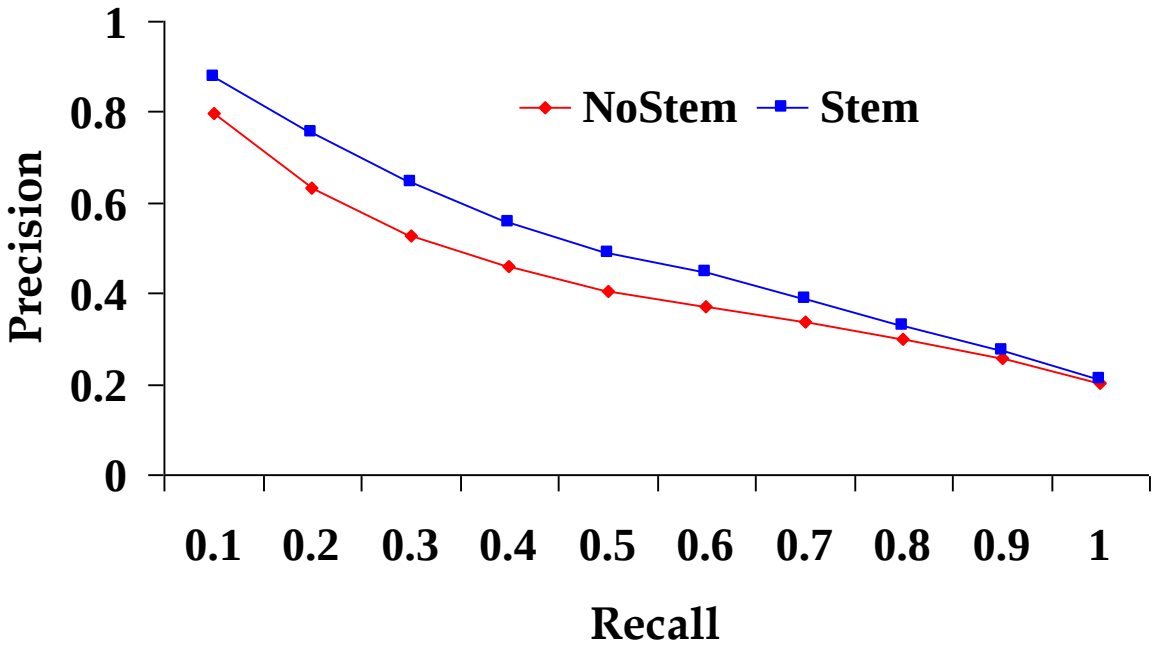
Average precision 52.0 %

Average Recall/Precision Curve

- Typically average performance over a large ***set*** of queries
- Compute average precision at each standard recall level across all queries
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus

Compare Two or More Systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance



Recall และ Precision

- แสดงให้เห็นถึงความสำเร็จที่สัมพันธ์กันของระบบที่สนองตอบต่อความต้องการต่าง ๆ ของผู้ใช้
- วิธีการวัดสามารถที่จะถูกตีความหมายโดยตรงในการแสดงดึงประสบการณ์ของผู้ใช้อีกด้วย

ดังนั้น

Precision Performance =0.2

Recall = 0.5

แสดงว่า

ผู้ใช้ได้รับเอกสารที่เกี่ยวข้องครึ่งหนึ่ง และจะต้องตรวจสอบเอกสารที่ไม่ถูกดึงออกมาทุกๆ 4 เอกสารเพราะมีเอกสารที่เกี่ยวข้องอยู่อีก

ปัญหาของวิธีการวัด Recall และ Precision

- วิธีการวัด Recall และ Precision ได้ถูกผูกเข้าด้วยกันตามปกติ (Normally Tied) ระหว่างกลุ่มเอกสารที่กำหนดให้มากับ Query Set ที่กำหนดภายในสิ่งแวดล้อมที่ไม่เปลี่ยนแปลง
- การเปลี่ยนแปลงวิธีการสร้างดัชนีหรือภาษาดัชนีหรือวิธีการค้นหา จะมีผลต่อการกระทำของระบบในแง่ของ Recall และ Precision

ปัญหาของวิธีการวัด Recall และ Precision

- การประเมินความเกี่ยวข้องของเอกสารกับข้อคำถามของผู้ใช้
- วิธีการวัด Recall และ Precision นั้นจะถูกนำไปใช้กับสิ่งแวดล้อมของผู้ใช้

ปัญหาของวิธีการวัด Recall และ Precision

- อิทธิพลของประเภทของข้อคำถามบนผลที่ได้จากการประเมินผล (Evaluation Outcome)

ประเภทของข้อคำถาม อย่างเช่น

1) ข้อคำถามประเภทที่มีหัวข้อเรื่องอย่างสั้น
2) ข้อคำถามประเภทที่ใช้ชื่อเรื่องทั้งหมด
3) ข้อคำถามประเภทที่ใช้ข้อความที่สมบูรณ์

- ข้อคำถามประเภทต่าง ๆ เหล่านี้ ได้ถูกใช้ในการที่จะสร้างข้อเรียกร้องในการค้นหา
- ถูกกำหนดโดยผู้ใช้ซึ่งเป็นสื่อกลางของวิธีการค้นหา ระบบควรที่จะถูกทดสอบโดยการใช้ข้อคำถามที่สามารถเป็นจริงผสมกับการสะท้อนให้เห็นประเภทของข้อคำถามที่ถูกส่งเข้าไปจริงในสถานการณ์ปฏิบัติ

ปัญหาของวิธีการวัด Recall และ Precision

- กำหนดระดับของความเกี่ยวข้องที่มีต่อเอกสารต่าง ๆ ของกลุ่มเอกสารที่กำหนดให้มา และการเลือกระดับของเอกสาร (Document Rank)

Other Evaluation Methods

- R- Precision
- The Harmonic Mean
- E-Measure
- F-Measure
- User-Oriented Measure

การแก้ไข โดยใช้ The E-Measure

เป็นการรวม Recall และ Precision ให้เป็นจำนวนเดียว โดยมีสูตรในการคำนวณดังนี้

$$E = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\frac{1}{R}}$$

$$\alpha = 1/(\beta^2 + 1)$$

โดย P = precision ,R = recall

ผลลัพธ์ของการคำนวณจะใช้สำหรับวัดความสัมพันธ์ที่สำคัญของ P และ R

ตัวอย่าง ถ้าผลลัพธ์จากการคำนวณ

E = 1 หมายความว่าผู้ใช้มีความสนใจใน precision และ recall เท่ากัน

E= ∞ หมายความว่าผู้ใช้ไม่สนใจเกี่ยวกับ precision

E= 0 หมายความว่าผู้ใช้ไม่สนใจเกี่ยวกับ recall

Fallout

- เป็นตัววัดซึ่งจะสะท้อนให้เห็นการกระทำสำหรับเอกสารที่ไม่เกี่ยวข้อง

$$\text{Fallout} = \frac{\text{จำนวนของเอกสารที่ไม่เกี่ยวข้องที่ถูกดึงออกมา}}{\text{จำนวนทั้งหมดของเอกสารที่ไม่เกี่ยวข้องในกลุ่มเอกสาร}}$$

Recall & Fallout

- ถ้า Recall ถูกแสดงในรูปแบบของความน่าจะเป็นของเอกสารหนึ่งที่ถูกดึงออกมาและมีความเกี่ยวข้องกับข้อเรียกร้อง
- Fallout ก็จะเป็นความน่าจะเป็นของเอกสารหนึ่งที่ถูกดึงออกมาและไม่เกี่ยวข้อง
- ดังนั้น ระบบค้นคืนสารสนเทศที่มีประสิทธิผลดี

จะมี Recall ที่สูงสุดและ Fallout ต่ำที่สุด

Generality Factor

- เป็นจำนวนเฉลี่ยของเอกสารที่เกี่ยวข้องที่มีอยู่ในกลุ่มเอกสารต่อหนึ่งข้อคำถาม
- ในสถานการณ์การดึงข้อมูลโดยปกติ, Recall, Precision และ Fallout จะขึ้นอยู่กับ Generality Factor

$$P = \frac{R.G}{(R.G) + F(1 - G)}$$

Contingency Table

	Relevant	Nonrelevant	
Retrieved	RETREL	RETNREL	RETREL+RETNREL
Not retrieved	NRETREL	NRETREL	NRETREL+ RETNREL
	RETREL	RETREL	RETREL+ RETNREL+
	+NRETREL	+NRETNREL	NRETREL RETNREL

Typical Retrieval Evaluation Measures

Symbol	Evaluation measure	Formula	Explanation
R	Recall	$\frac{\text{RETREL}}{\text{RETREL} + \text{NRETREL}}$	Proportion of relevant actually retrieved
P	Precision	$\frac{\text{RETREL}}{\text{RETREL} + \text{RETNREL}}$	Proportion of retrieved actually relevant
F	Fallout	$\frac{\text{RETNREL}}{\text{RETNREL} + \text{NRETNREL}}$	Proportion of nonrelevant actually retrieved
G	Generality	$\frac{\text{RETREL} + \text{NRETREL}}{\text{RETREL} + \text{NRETREL} + \text{NRETREL} + \text{NRETNREL}}$	Proportion of relevant per query

ตัวอย่าง การประเมินการค้นคืนสารสนเทศ

- สมมุติว่าจะทำการประเมินระบบสามระบบคือ
  - “Bear”
  - “Cardinal”
  - “Wolf”
- โดยใช้ความเกี่ยวข้องของสารสนเทศและระดับ (relevance information and rankings)
- สำหรับ สาม ข้อเรียกร้อง(three queries) กระทำบน Excel spreadsheet

สิ่งสำคัญที่ต้องการในการประเมินสารสนเทศประกอบด้วย

- การรวบรวมข้อมูลที่ใช้สำหรับทดสอบ
- ระบบค้นคืนสารสนเทศ(IR systems) ที่สามารถให้ผลการจัดลำดับที่ตอบสนองต่อข้อสอบถาม(queries)
- ข้อสอบถามที่ต้องการทดสอบ
- การประเมินผลลัพธ์และความเกี่ยวพันธ์ที่ประเมินจากข้อสอบถาม (RELEVANCE JUDGEMENTS on the Queries) ซึ่งส่วนมากจะได้ผลลัพธ์เป็นเลขไบนารี (relevant or not relevant) หรือใช้ two scales

Relevance	Judgements		
Document	Query 0	Query 1	Query 2
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	1	0	0
6	0	0	0
7	0	0	0
8	0	1	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0
13	0	0	0
14	0	0	1
15	0	0	1
16	0	0	1
17	0	0	1
18	0	0	1
19	0	0	0
20	0	0	0
21	0	0	0
22	0	0	0
23	0	0	0
24	0	0	0
25	0	0	0

Rank	Query 0		Rankings			
	Bear	Rel	**Cardinal**	Rel	**Wolf**	Rel
1	5		7		11	
2	2		16		2	
3	1		9		10	
4	15		18		21	
5	9		4		1	
6	19		17		5	
7	3		8		22	
8	6		5		9	
9	18		11		20	
10	4		15		3	
11	12		12		23	
12	8		10		4	
13	11		6		19	
14	17		2		6	
15	7		19		15	
16	20		20		25	
17	10		21		18	
18	21		25		16	
19	16		22		7	
20	23		1		8	
21	13		23		17	
22	22		13		12	
23	24		24		13	
24	25		3		14	
25	14		14		24	

	Query 1		Rankings			
Rank	**Bear**	Rel	**Cardinal**	Rel	**Wolf**	Rel
1	12		16		7	
2	14		23		23	
3	8		3		3	
4	23		13		13	
5	9		4		4	
6	5		14		8	
7	20		17		17	
8	21		11		25	
9	4		7		16	
10	19		21		21	
11	3		18		18	
12	10		10		22	
13	7		19		19	
14	18		22		10	
15	11		20		20	
16	17		1		5	
17	24		2		2	
18	13		12		12	
19	2		15		15	
20	16		5		1	
21	6		8		14	
22	22		6		24	
23	15		9		9	
24	25		25		11	
25	1		24		6	

	Query 2		Rankings			
Rank	Bear	Rel	Cardinal	Rel	Wolf	Rel
1	1		5		23	
2	18		24		9	
3	11		13		25	
4	16		19		18	
5	19		17		1	
6	10		12		2	
7	23		22		11	
8	21		4		10	
9	3		1		19	
10	6		8		16	
11	22		11		12	
12	12		3		22	
13	17		16		15	
14	2		23		8	
15	13		25		24	
16	24		18		14	
17	25		2		7	
18	14		20		13	
19	5		14		17	
20	4		10		20	
21	7		6		21	
22	9		21		4	
23	15		15		5	
24	20		7		6	
25	8		9		3	

Rank	Query 0		Rankings			
	Bear	Rel	**Cardinal**	Rel	**Wolf**	Rel
1	5	1	7	0	11	0
2	2	2	16	0	2	1
3	1	3	9	0	10	1
4	15	3	18	0	21	1
5	9	3	4	1	1	2
6	19	3	17	1	5	3
7	3	4	8	1	22	3
8	6	4	5	2	9	3
9	18	4	11	2	20	3
10	4	5	15	2	3	4
11	12	5	12	2	23	4
12	8	5	10	2	4	5
13	11	5	6	2	19	5
14	17	5	2	3	6	5
15	7	5	19	3	15	5
16	20	5	20	3	25	5
17	10	5	21	3	18	5
18	21	5	25	3	16	5
19	16	5	22	3	7	5
20	23	5	1	4	8	5
21	13	5	23	4	17	5
22	22	5	13	4	12	5
23	24	5	24	4	13	5
24	25	5	3	5	14	5
25	14	5	14	5	24	5

	Query 1		Rankings		
Rank	**Bear**	Rel	**Cardinal**	Rel	**Wolf**
1	12	1	16	0	7
2	14	1	23	0	23
3	8	2	3	0	3
4	23	2	13	0	13
5	9	3	4	0	4
6	5	3	14	0	8
7	20	3	17	0	17
8	21	3	11	1	25
9	4	3	7	1	16
10	19	3	21	1	21
11	3	3	18	1	18
12	10	4	10	2	22
13	7	4	19	2	19
14	18	4	22	2	10
15	11	5	20	2	20
16	17	5	1	2	5
17	24	5	2	2	2
18	13	5	12	3	12
19	2	5	15	3	15
20	16	5	5	3	1
21	6	5	8	4	14
22	22	5	6	4	24
23	15	5	9	5	9
24	25	5	25	5	11
25	1	5	24	5	6

	Query 2		Rankings			
Rank	Bear	Rel	Cardinal	Rel	Wolf	Rel
1	1	0	5	0	23	0
2	18	1	24	0	9	0
3	11	1	13	0	25	0
4	16	2	19	0	18	1
5	19	2	17	1	1	1
6	10	2	12	1	2	1
7	23	2	22	1	11	1
8	21	2	4	1	10	1
9	3	2	1	1	19	1
10	6	2	8	1	16	2
11	22	2	11	1	12	2
12	12	2	3	1	22	2
13	17	3	16	2	15	3
14	2	3	23	2	8	3
15	13	3	25	2	24	3
16	24	3	18	3	14	4
17	25	3	2	3	7	4
18	14	4	20	3	13	4
19	5	4	14	4	17	5
20	4	4	10	4	20	5
21	7	4	6	4	21	5
22	9	4	21	4	4	5
23	15	5	15	5	5	5
24	20	5	7	5	6	5
25	8	5	9	5	3	5

คำนวณหาค่า **recall** และ **precision** จากสูตร

$$\text{Recall} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Rel in Collection}\,|}$$

$$\text{Precision} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Retrieved}\,|}$$

By systems… for Bear…

	Query 0		Query 1		Query 2	
Bear	Recall	Prec	Recall	Prec	Recall	Prec
1	0.20	1.00	0.20	1.00	0.00	0.00
2	0.40	1.00	0.20	0.50	0.20	0.50
3	0.60	1.00	0.40	0.67	0.20	0.33
4	0.60	0.75	0.40	0.50	0.40	0.50
5	0.60	0.60	0.60	0.60	0.40	0.40
6	0.60	0.50	0.60	0.50	0.40	0.33
7	0.80	0.57	0.60	0.43	0.40	0.29
8	0.80	0.50	0.60	0.38	0.40	0.25
9	0.80	0.44	0.60	0.33	0.40	0.22
10	1.00	0.50	0.60	0.30	0.40	0.20
11	1.00	0.45	0.60	0.27	0.40	0.18
12	1.00	0.42	0.80	0.33	0.40	0.17
13	1.00	0.38	0.80	0.31	0.60	0.23
14	1.00	0.36	0.80	0.29	0.60	0.21
15	1.00	0.33	1.00	0.33	0.60	0.20
16	1.00	0.31	1.00	0.31	0.60	0.19
17	1.00	0.29	1.00	0.29	0.60	0.18
18	1.00	0.28	1.00	0.28	0.80	0.22
19	1.00	0.26	1.00	0.26	0.80	0.21
20	1.00	0.25	1.00	0.25	0.80	0.20
21	1.00	0.24	1.00	0.24	0.80	0.19
22	1.00	0.23	1.00	0.23	0.80	0.18
23	1.00	0.22	1.00	0.22	1.00	0.22
24	1.00	0.21	1.00	0.21	1.00	0.21
25	1.00	0.20	1.00	0.20	1.00	0.20

For Cardinal

	Query 0		Query 1		Query 2	
Cardinal	Recall	Prec	Recall	Prec	Recall	Prec
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.20	0.20	0.00	0.00	0.20	0.20
6	0.20	0.17	0.00	0.00	0.20	0.17
7	0.20	0.14	0.00	0.00	0.20	0.14
8	0.40	0.25	0.20	0.13	0.20	0.13
9	0.40	0.22	0.20	0.11	0.20	0.11
10	0.40	0.20	0.20	0.10	0.20	0.10
11	0.40	0.18	0.20	0.09	0.20	0.09
12	0.40	0.17	0.40	0.17	0.20	0.08
13	0.40	0.15	0.40	0.15	0.40	0.15
14	0.60	0.21	0.40	0.14	0.40	0.14
15	0.60	0.20	0.40	0.13	0.40	0.13
16	0.60	0.19	0.40	0.13	0.60	0.19
17	0.60	0.18	0.40	0.12	0.60	0.18
18	0.60	0.17	0.60	0.17	0.60	0.17
19	0.60	0.16	0.60	0.16	0.80	0.21
20	0.80	0.20	0.60	0.15	0.80	0.20
21	0.80	0.19	0.80	0.19	0.80	0.19
22	0.80	0.18	0.80	0.18	0.80	0.18
23	0.80	0.17	1.00	0.22	1.00	0.22
24	1.00	0.21	1.00	0.21	1.00	0.21
25	1.00	0.20	1.00	0.20	1.00	0.20

For Wolf

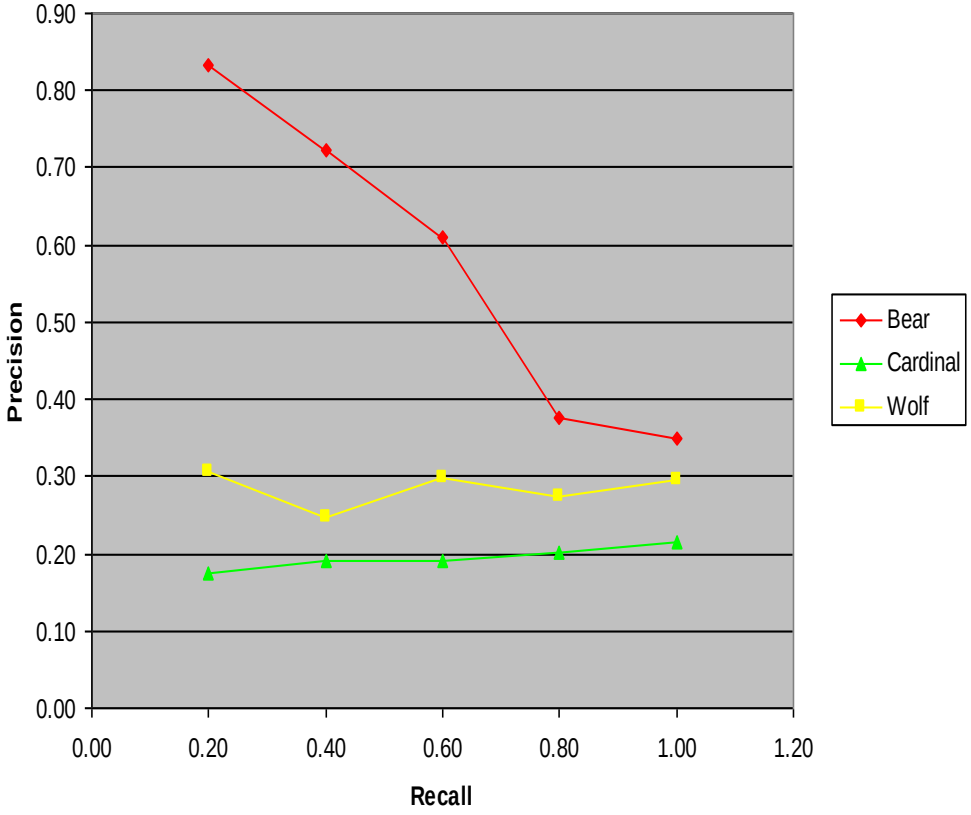
	Query 0		Query 1		Query 2	
Wolf	Recall	Prec	Recall	Prec	Recall	Prec
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.20	0.50	0.00	0.00	0.00	0.00
3	0.20	0.33	0.00	0.00	0.00	0.00
4	0.20	0.25	0.00	0.00	0.20	0.25
5	0.40	0.40	0.00	0.00	0.20	0.20
6	0.60	0.50	0.20	0.17	0.20	0.17
7	0.60	0.43	0.20	0.14	0.20	0.14
8	0.60	0.38	0.20	0.13	0.20	0.13
9	0.60	0.33	0.20	0.11	0.20	0.11
10	0.80	0.40	0.20	0.10	0.40	0.20
11	0.80	0.36	0.20	0.09	0.40	0.18
12	1.00	0.42	0.20	0.08	0.40	0.17
13	1.00	0.38	0.20	0.08	0.60	0.23
14	1.00	0.36	0.40	0.14	0.60	0.21
15	1.00	0.33	0.40	0.13	0.60	0.20
16	1.00	0.31	0.40	0.13	0.80	0.25
17	1.00	0.29	0.40	0.12	0.80	0.24
18	1.00	0.28	0.60	0.17	0.80	0.22
19	1.00	0.26	0.60	0.16	1.00	0.26
20	1.00	0.25	0.60	0.15	1.00	0.25
21	1.00	0.24	0.60	0.14	1.00	0.24
22	1.00	0.23	0.60	0.14	1.00	0.23
23	1.00	0.22	0.80	0.17	1.00	0.22
24	1.00	0.21	1.00	0.21	1.00	0.21
25	1.00	0.20	1.00	0.20	1.00	0.20

Averages for all systems

- ทำการหาค่าเฉลี่ยของprecision ที่ระดับของ recall ที่กำหนด ของระบบทั้งหมดกำหนดไว้ 5 ระดับได้แก่
  - 20% หรือ 0.20
  - 40% หรือ 0.40
  - 60% หรือ 0.60
  - 80% หรือ 0.80
  - 100% หรือ 1.00
- การคำนวณแต่ละข้อสอบถามนั้น หาจากค่าของ precision ในลำดับที่สูงที่สุด(HIGHEST RANKED precision value) ที่จับคู่กับค่าของ Recall ในระดับที่ยอมรับ

Averages for all systems

	Bear	Cardinal	Wolf
0.20	0.83	0.18	0.31
0.40	0.72	0.19	0.25
0.60	0.61	0.19	0.30
0.80	0.38	0.20	0.27
1.00	0.35	0.21	0.30



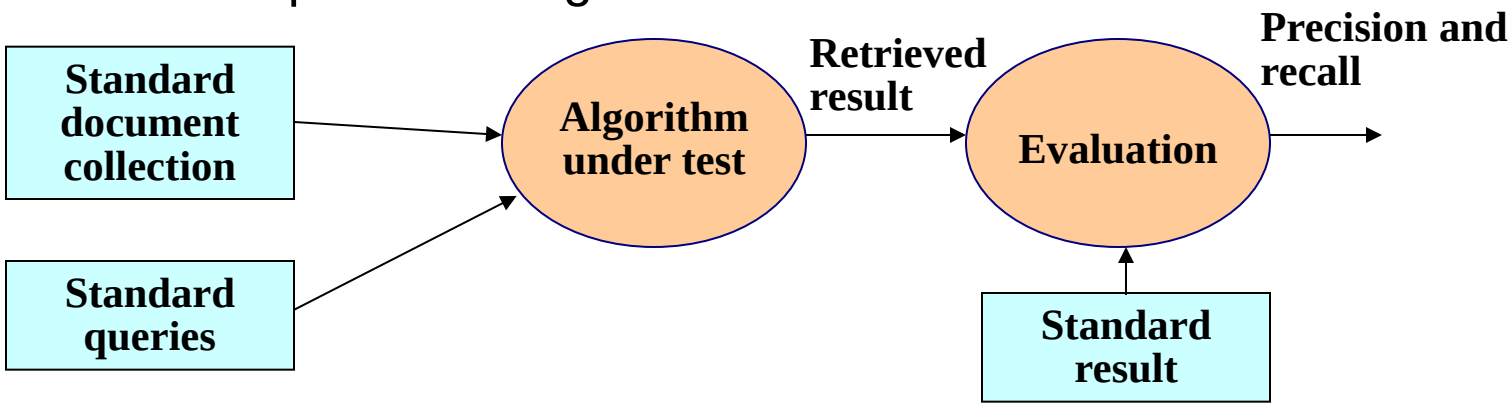
Benchmarking

Experimental Setup for Benchmarking

- ***Analytical*** performance evaluation is difficult for document retrieval systems because many characteristics such as relevance, distribution of words, etc., are difficult to describe with mathematical precision
- Performance is measured by ***benchmarking***. That is, the retrieval effectiveness of a system is evaluated on a *given set of documents*, *queries*, and *relevance judgments*
- Performance data is valid only for the environment under which the system is evaluated

Benchmarks

- A benchmark collection contains:
  - A set of standard documents and queries/topics
  - A list of relevant documents for each query
- Standard collections for traditional IR:
  - Smart collection: ftp://ftp.cs.cornell.edu/pub/smart
  - TREC: http://trec.nist.gov/



Benchmarking Problems

- Performance data is valid only for a particular benchmark
- Building a benchmark corpus is a difficult task
- Benchmark web corpora are just starting to be developed
- Benchmark foreign-language corpora are just starting to be developed

Early Test Collections

- Previous experiments were based on the SMART collection which is fairly small (ftp://ftp.cs.cornell.edu/pub/smart)

Collection Name	Number Of Documents	Number Of Queries	Raw Size (Mbytes)
CACM	3,204	64	1.5
CISI	1,460	112	1.3
CRAN	1,400	225	1.6
MED	1,033	30	1.1
TIME	425	83	1.5

- Different researchers used different test collections and evaluation techniques

The TREC Benchmark

- TREC: Text REtrieval Conference (http://trec.nist.gov/) Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA)
- Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA
- Participants are given parts of a standard set of documents and TOPICS (from which queries have to be derived) in different stages for training and testing
- Participants submit the P/R values for the final document and query corpus and present their results at the conference

The TREC Objectives

- Provide a common ground for comparing different IR techniques
  - Same set of documents and queries, and same evaluation method
- Sharing of resources and experiences in developing the benchmark
  - With major sponsorship from government to develop large benchmark collections
- Encourage participation from industry and academia
- Development of new evaluation techniques, particularly for new applications
  - Retrieval, routing/filtering, non-English collection, web-based collection, question answering

TREC Advantages

- Large scale (compared to a few MB in the SMART Collection)
- Relevance judgments provided
- Under continuous development with support from the U.S. Government
- Wide participation:
  - TREC 1: 28 papers 360 pages.
  - TREC 4: 37 papers 560 pages.
  - TREC 7: 61 papers 600 pages.
  - TREC 8: 74 papers

TREC Tasks

- New tasks added after TREC 5 - Interactive, multilingual, natural language, multiple database merging, filtering, very large corpus (20 GB, 7.5 million documents), question answering

Cross Language Evaluation Forum (CLEF)

- European counterpart of TREC
- Promoting research and development in cross language information retrieval by
  - Providing an infrastructure for the testing and evaluation of information retrieval system operating on European languages in both monolingual and cross language contexts
  - Creating test suites of reusable data which can be employed by system developers for benchmarking purpose
- www.clef-campaign.org

Characteristics of the TREC Collection

- Both long and short documents (from a few hundred to over one thousand unique terms in a document).
- Test documents consist of:

WSJ	Wall Street Journal articles (1986-1992)	550 M
AP	Associate Press Newswire (1989)	514 M
ZIFF	Computer Select Disks (Ziff-Davis Publishing)	493 M
FR	Federal Register	469 M
DOE	Abstracts from Department of Energy reports	190 M

More Details on Document Collections

- Volume 1 (Mar 1994) - Wall Street Journal (1987, 1988, 1989), Federal Register (1989), Associated Press (1989), Department of Energy abstracts, and Information from the Computer Select disks (1989, 1990)
- Volume 2 (Mar 1994) - Wall Street Journal (1990, 1991, 1992), the Federal Register (1988), Associated Press (1988) and Information from the Computer Select disks (1989, 1990)
- Volume 3 (Mar 1994) - San Jose Mercury News (1991), the Associated Press (1990), U.S. Patents (1983-1991), and Information from the Computer Select disks (1991, 1992)

More Details on Document Collections

- Volume 4 (May 1996) - Financial Times Limited (1991, 1992, 1993, 1994), the Congressional Record of the 103rd Congress (1993), and the Federal Register (1994)

- Volume 5 (Apr 1997) - Foreign Broadcast Information Service (1996) and the Los Angeles Times (1989, 1990)

TREC Properties

- Both documents and queries contain many different kinds of information (fields)
- Generation of the formal queries (Boolean, Vector Space, etc.) is the responsibility of the system
  - A system may be very good at querying and ranking, but if it generates poor queries from the topic, its final P/R would be poor

Cystic Fibrosis (CF) Collection

- 1,239 abstracts of medical journal articles on CF
- 100 information requests (queries) in the form of complete English questions
- Relevant documents determined and rated by 4 separate medical experts on 0-2 scale:
  - 0: Not relevant
  - 1: Marginally relevant
  - 2: Highly relevant

CF Document Fields

- MEDLINE access number
- Author
- Title
- Source
- Major subjects
- Minor subjects
- Abstract (or extract)
- References to other documents
- Citations to this document

Sample CF Document

AN 74154352
AU Burnell-R-H. Robertson-E-F.
TI Cystic fibrosis in a patient with Kartagener syndrome.
SO Am-J-Dis-Child. 1974 May. 127(5). P 746-7.
MJ CYSTIC-FIBROSIS: co. KARTAGENER-TRIAD: co.
MN CASE-REPORT. CHLORIDES: an. HUMAN. INFANT. LUNG: ra. MALE.
SITUS-INVERSUS: co, ra. SODIUM: an. SWEAT: an.
AB A patient exhibited the features of both Kartagener syndrome and cystic fibrosis. At most, to the authors' knowledge, this represents the third such report of the combination. Cystic fibrosis should be excluded before a diagnosis of Kartagener syndrome is made.
RF 001 KARTAGENER M BEITR KLIN TUBERK 83 489 933
002 SCHWARZ V ARCH DIS CHILD 43 695 968
003 MACE JW CLIN PEDIATR 10 285 971
…
CT 1 BOCHKOVA DN GENETIKA (SOVIET GENETICS) 11 154 975
2 WOOD RE AM REV RESPIR DIS 113 833 976
3 MOSSBERG B MT SINAI J MED 44 837 977

References

- Modern Information Retrieval, by Beaza-Yates and Ribeiro-Neto, 1999
- James Allen, University of Massachusetts Amherst
- Raymond J. Mooney, University of Texas
- Christof Monz and Maarten de Rijke, University of Amsterdam