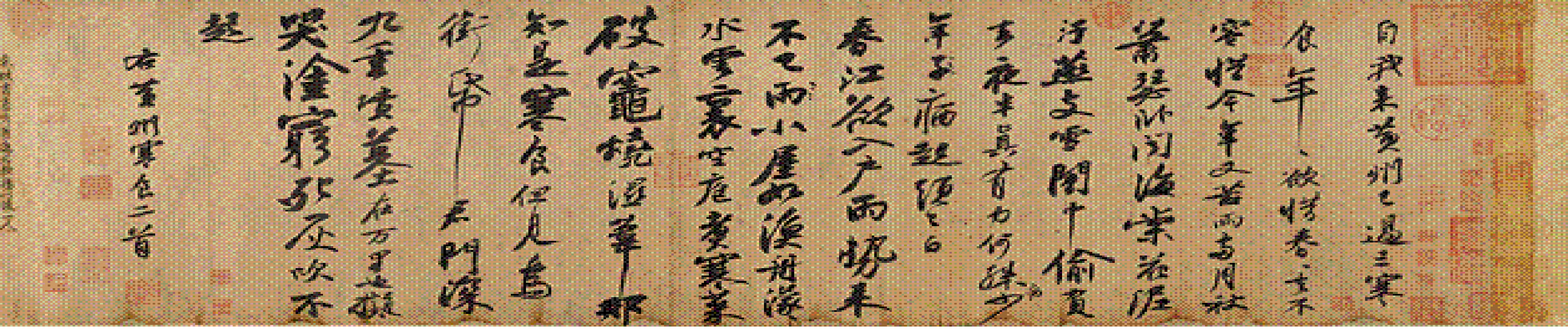


信息检索
Information Retrieval

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA：胡锦毅

Email:hu-jy21@mails.tsinghua.edu.cn

郑重声明

- 此课件仅供选修清华大学计算机系本科生课《信息检索》（课号：40240372）的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

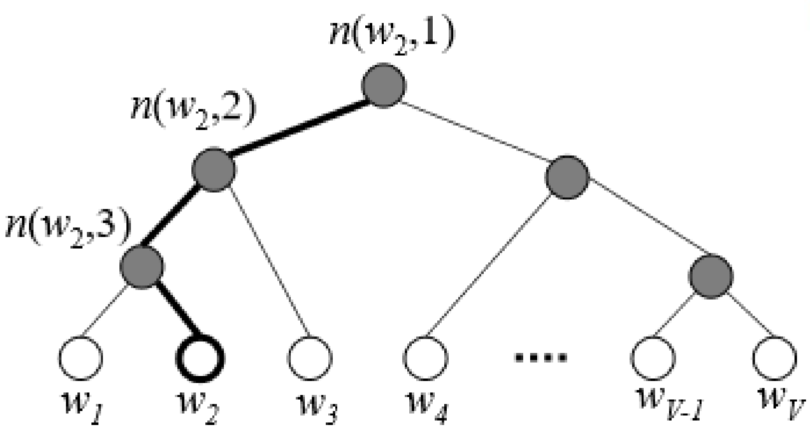
Word2Vec(续)

Hierarchical softmax(Morin and Bengio)

More precisely, each word w can be reached by an appropriate path from the root of the tree. Let $n(w,j)$ be the j-th node on the path from the root to w, and let $L(w)$ be the length of this path, so $n(w,1) = \text{root}$ and $n(w,L(w)) = w$. In addition, for any inner node n, let $\text{ch}(n)$ be an arbitrary fixed child of n and let $[\![x]\!]$ be 1 if x is true and -1 otherwise. Then the hierarchical softmax defines $p(w_O|w_I)$ as follows:

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma\left([\![n(w,j+1) = \text{ch}(n(w,j))]\!] \cdot {v'_{n(w,j)}}^\top v_{w_I}\right)$$

where $\sigma(x) = 1/(1+\exp(-x))$. It can be verified that $\sum_{w=1}^{W} p(w|w_I) = 1$.



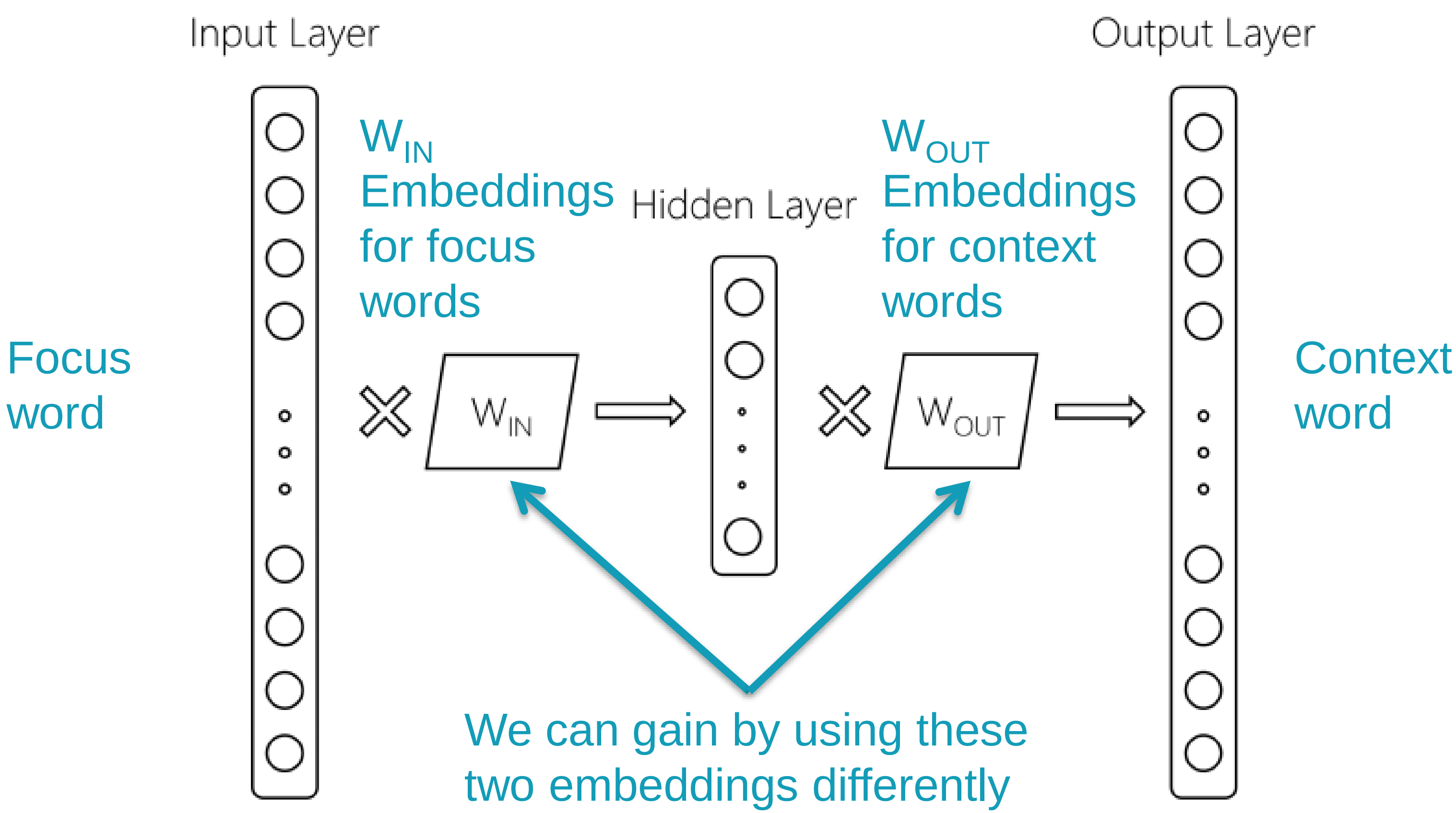
$$\begin{aligned} p(w_2 = w_O) &= p(n(w_2,1),\text{left}) \cdot p(n(w_2,2),\text{left}) \cdot p(n(w_2,3),\text{right}) \\ &= \sigma\left({\mathbf{v}'_{n(w_2,1)}}^T \mathbf{h}\right) \cdot \sigma\left({\mathbf{v}'_{n(w_2,2)}}^T \mathbf{h}\right) \cdot \sigma\left(-{\mathbf{v}'_{n(w_2,3)}}^T \mathbf{h}\right) \end{aligned}$$

$$O(V) \rightarrow O(\log_2 V)$$

https://medium.com/@ameyyadav/hierarchical-softmax-as-output-activation-function-in-neural-network-part-2-e6434131e203

Using 2 word embeddings

word2vec model with 1 word of context



Using 2 word embeddings

yale		seahawks	
IN-IN	IN-OUT	IN-IN	IN-OUT
yale	yale	seahawks	seahawks
harvard	faculty	49ers	highlights
nyu	alumni	broncos	jerseys
cornell	orientation	packers	tshirts
tulane	haven	nfl	seattle
tufts	graduate	steelers	hats

Dual Embedding Space Model (DESM)

- Simple model
- A document is represented by the centroid of its word vectors

$$\overline{\mathbf{D}} = \frac{1}{|D|} \sum_{\mathbf{d}_j \in D} \frac{\mathbf{d}_j}{\|\mathbf{d}_j\|}$$

- Query-document similarity is average over query words of cosine similarity

$$DESM(Q, D) = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{\mathbf{q}_i^T \overline{\mathbf{D}}}{\|\mathbf{q}_i\| \|\overline{\mathbf{D}}\|}$$

Dual Embedding Space Model (DESM)

- What works best is to use the OUT vectors for the document and the IN vectors for the query

$$DESM_{IN-OUT}(Q,D) = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{q_{IN,i}^{T} \overline{D_{OUT}}}{\|q_{IN,i}\| \|\overline{D_{OUT}}\|}$$

- This way similarity measures *aboutness* – words that appear with this word – which is more useful in this context than *(distributional) semantic similarity*

Experiments

- Train word2vec from either
  - 600 million Bing queries
  - 342 million web document sentences
- Test on 7,741 randomly sampled Bing queries
  - 5 level eval (Perfect, Excellent, Good, Fair, Bad)
- Two approaches
  1. Use DESM model to rerank top results from BM25
  2. Use DESM alone or a mixture model of it and BM25

$$MM(Q,D) = \alpha DESM(Q,D) + (1-\alpha)BM25(Q,D)$$

$$\alpha \in \mathbb{R}, 0 \leq \alpha \leq 1$$

Results – reranking *k*-best list

	Explicitly Judged Test Set		
	NDCG@1	NDCG@3	NDCG@10
BM25	23.69	29.14	44.77
LSA	22.41*	28.25*	44.24*
DESM (IN-IN, trained on body text)	23.59	29.59	45.51*
DESM (IN-IN, trained on queries)	23.75	29.72	46.36*
DESM (IN-OUT, trained on body text)	24.06	30.32*	46.57*
DESM (IN-OUT, trained on queries)	**25.02***	**31.14***	**47.89***

Pretty decent gains – e.g., 2% for NDCG@3

Gains are bigger for model trained on queries than docs

第五章 检索评价

5.1 Evaluation of Retrieval Efficiency and Effectiveness

- Effectiveness

There are many retrieval models/ algorithms/ systems, which one is the best?

What is the best component for:

Ranking function (dot-product, cosine, …)

Term selection (stopword removal, stemming…)

Term weighting (TF, TF-IDF,…)

"capable of retrieving what they want and of rejecting what they do not want."

- Efficiency

the user effort, the time, and the cost necessary to perform the retrieval task

5.2 Evaluation of Retrieval: Effectiveness

- The interpretation of relevance
  not only the contents of a document but also the state of knowledge of the user at the time of the search.

Relevancy is not typically binary but continuous.

Even if relevancy is binary, it can be a difficult judgment to make：

Subjective: Depends upon a specific user's judgment. 因人而异

Situational: Relates to user's current needs.因需而异

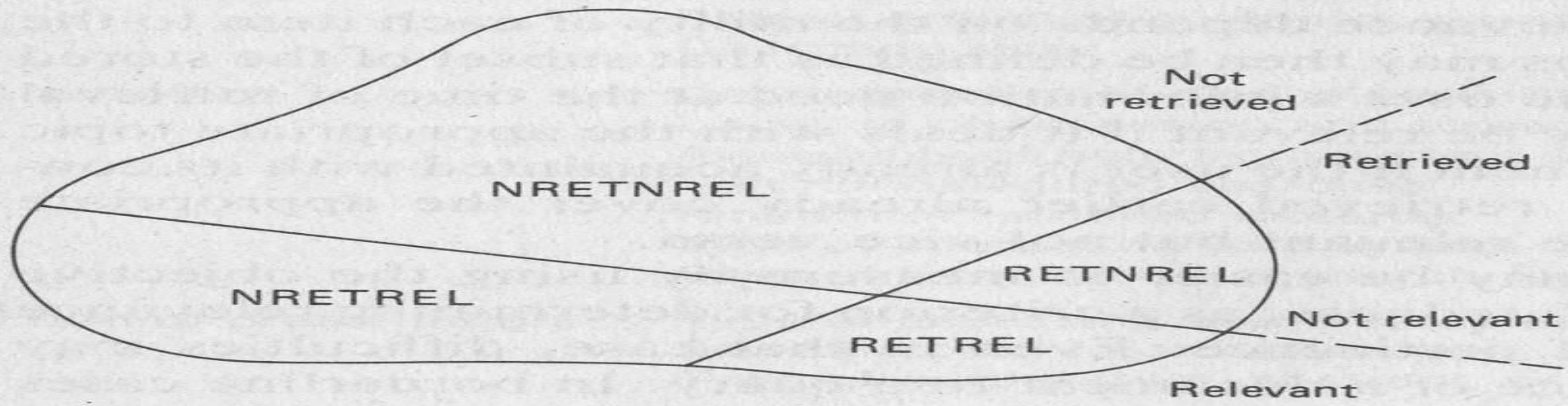
Dynamic: Changes over time. 因时而异

5.2 Evaluation of Retrieval: Effectiveness

- Human Labeled Corpora (Gold Standard)

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.

5.2 Evaluation of Retrieval: Effectiveness



RETREL : number relevant and retrieved
RETNREL : number not relevant and retrieved
NRETREL : number relevant and not retrieved
NRETNREL: number not relevant and not retrieved

Figure 5-1 Partition of collection.

$$R = \frac{number-of-items-retrieved-and-relevant}{total-relevant-in-collection}$$

$$P = \frac{number-of-items-retrieved-and-relevant}{total-retrieved}$$

	retrieved	not retrieved
Irrelevant (non-target)	False alarm	correct
Relevant (targets)	correct	Missed detection

5.2 Evaluation of Retrieval: Effectiveness

	Relevant	Non-relevant	Total
Retrieved	A	B	A+B
Not retrieved	C	D	C+D
Total	A+C	B+D	A+B+C+D

Recall: $\frac{A}{A+C}$ – proportion of retrieved items amongst the relevant items

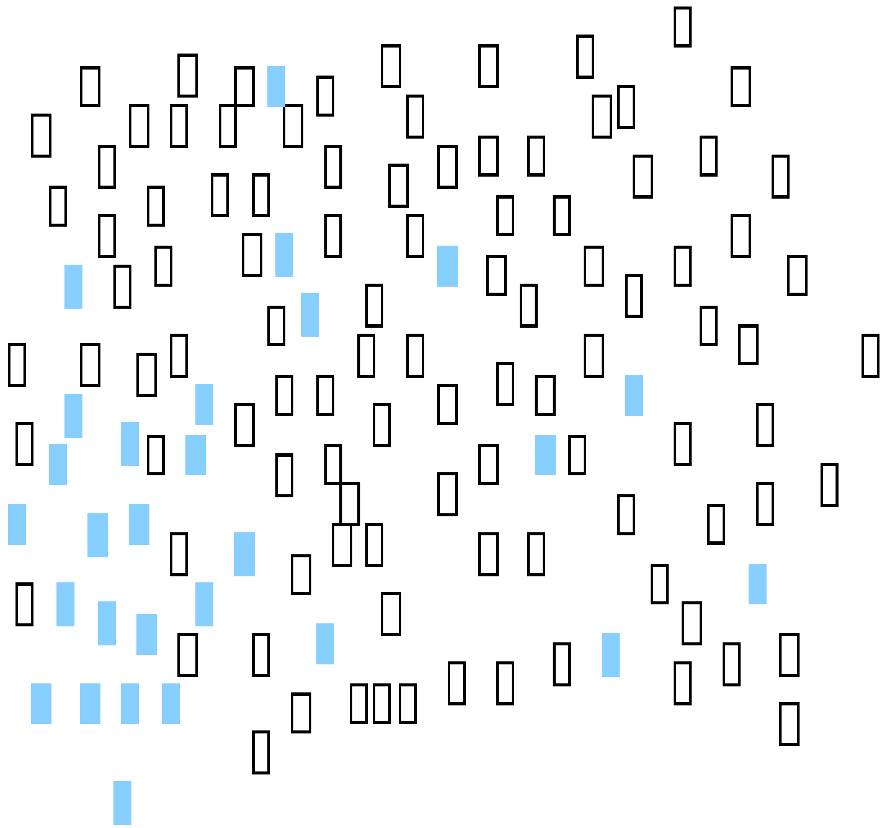
Precision: $\frac{A}{A+B}$ – proportion of relevant items amongst retrieved items

Accuracy: $\frac{A+D}{A+B+C+D}$ – proportion of correctly classified items as relevant/irrelevant

Recall: [0..1]; Precision: [0..1]; Accuracy: [0..1]

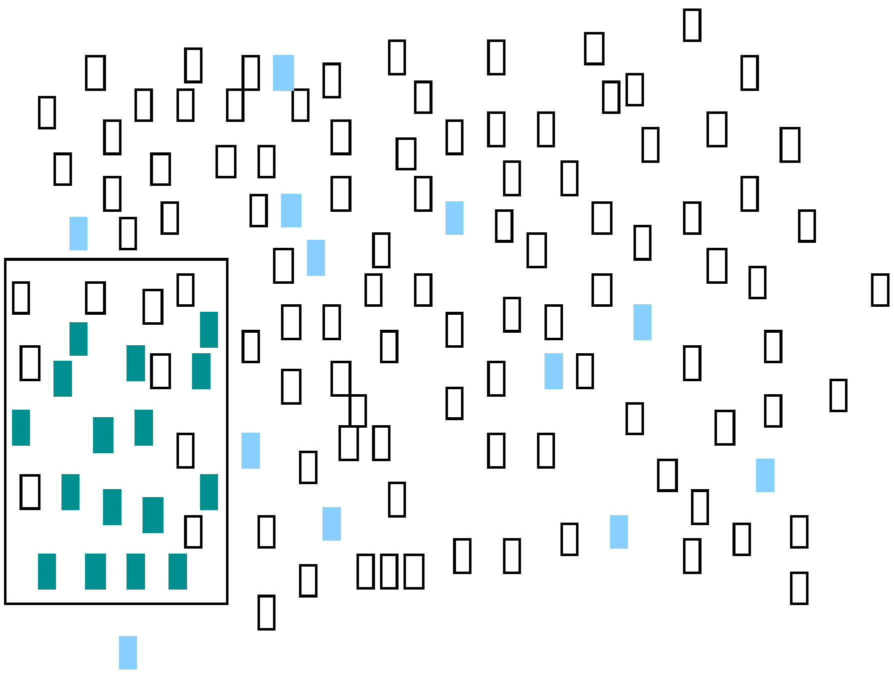
5.2 Evaluation of Retrieval: Effectiveness

- All documents: A+B+C+D = 130
- Relevant documents for a given query: A+C = 28



5.2 Evaluation of Retrieval: Effectiveness

- System 1 retrieves 25 items: $(A+B)_1 = 25$
- Relevant and retrieved items: $A_1 = 16$



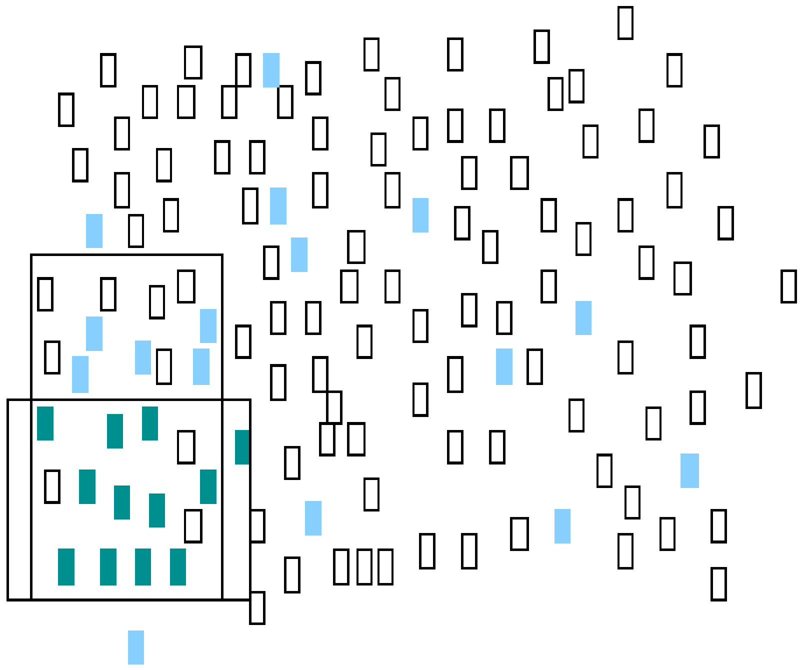
$$R_1 = \frac{A_1}{A+C} = \frac{16}{28} = .57$$

$$P_1 = \frac{A_1}{(A+B)_1} = \frac{16}{25} = .64$$

$$A_1 = \frac{A_1+D_1}{A+B+C+D} = \frac{16+93}{130} = .84$$

5.2 Evaluation of Retrieval: Effectiveness

- System B retrieves set $(A+B)_2$ = 15 items
- A_2 = 12



$$R_2 = \frac{12}{28} = .43$$

$$P_2 = \frac{12}{15} = .8$$

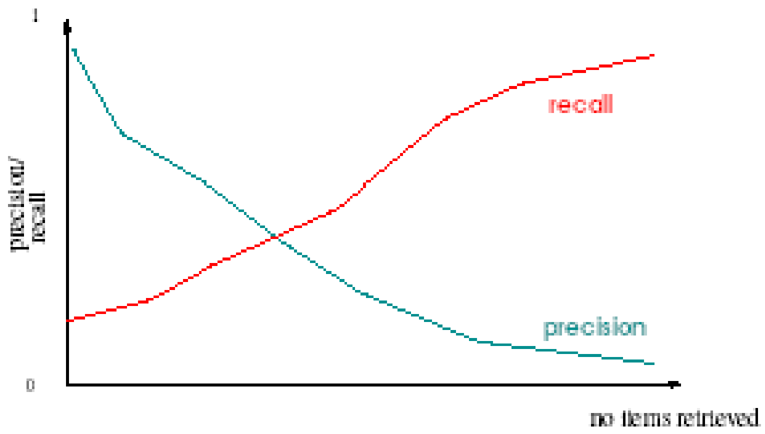
$$A_2 = \frac{12+99}{130} = .85$$

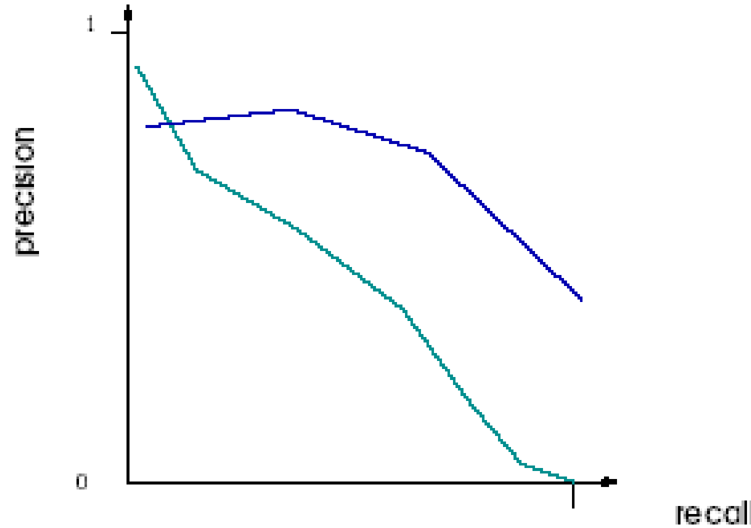
5.2 Evaluation of Retrieval: Effectiveness

- In general, one wants good precision and good recall
- But there is an inverse relationship between these

- The recall will increase as the number of retrieved documents increase; at the same time, the precision is likely to decrease.

5.2 Evaluation of Retrieval: Effectiveness



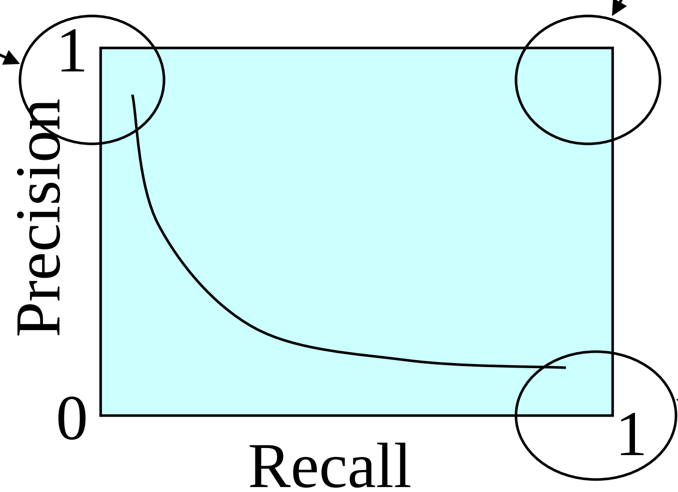


- Plotting precision and recall (versus no. of documents retrieved) shows inverse relationship between precison and recall
- Precision/recall cross-over as quality measure
- Plotting precision versus recall gives recall-precision curve
- Area under normalised recall-precision curve as quality measure

5.2 Evaluation of Retrieval: Effectiveness

Returns relevant documents but misses many useful ones too

The ideal



Returns most relevant documents but includes lots of junk

Trade-off between Recall and Precision:
Precision and Recall are inverse proportional

5.2 Evaluation of Retrieval: Effectiveness

- Prefer high recall or high precision?

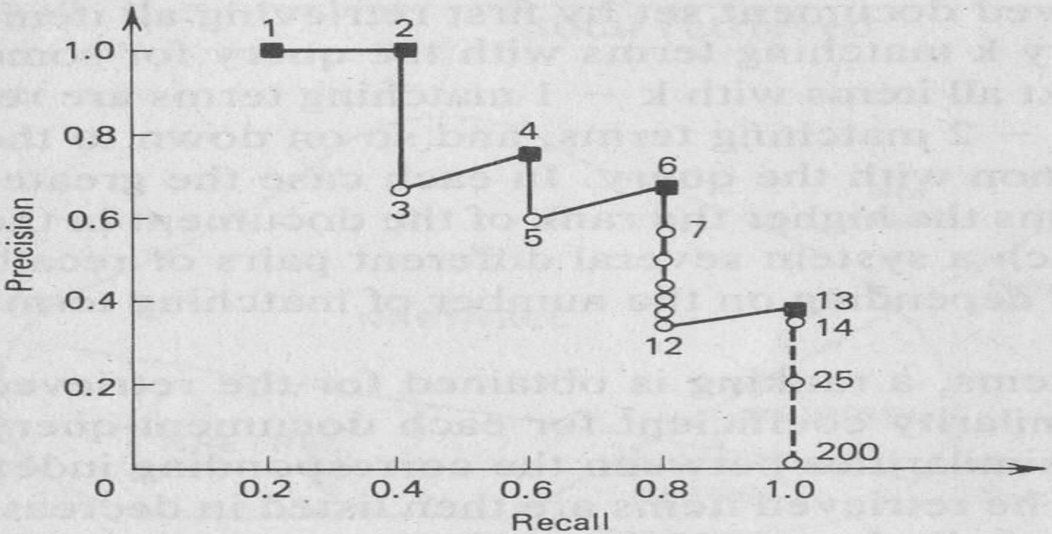
Precision-critical task	Recall-critical task
Little time available	Time matters less
A small set of relevant documents answers the information need	One cannot afford to miss a single document
Example: web search for factual information	Example: patent search

- The recall measurement requires information of the total number of relevant documents in the collection with respect to each query.

n	Document number (x = relevant)		Recall	Precision
1	588	x	0.2	1.0
2	589	x	0.4	1.0
3	576		0.4	0.67
4	590	x	0.6	0.75
5	986		0.6	0.60
6	592	x	0.8	0.67
7	984		0.8	0.57
8	988		0.8	0.50
9	578		0.8	0.44
10	985		0.8	0.40
11	103		0.8	0.36
12	591		0.8	0.33
13	772	x	1.0	0.38
14	990		1.0	0.36

Recall-precision after retrieval of n documents

(a)



(b)

Figure 5-2 Display of recall and precision results for a sample query. (Collection consists of 200 documents in aerodynamics.) (a) Output ranking of documents in decreasing query-document similarity order and computation of recall and precision values for a single query. (b) Graph of precision versus recall for sample query of Fig. 5-2a.

5.2 Evaluation of Retrieval: Effectiveness

- The average precision P at the recall level R

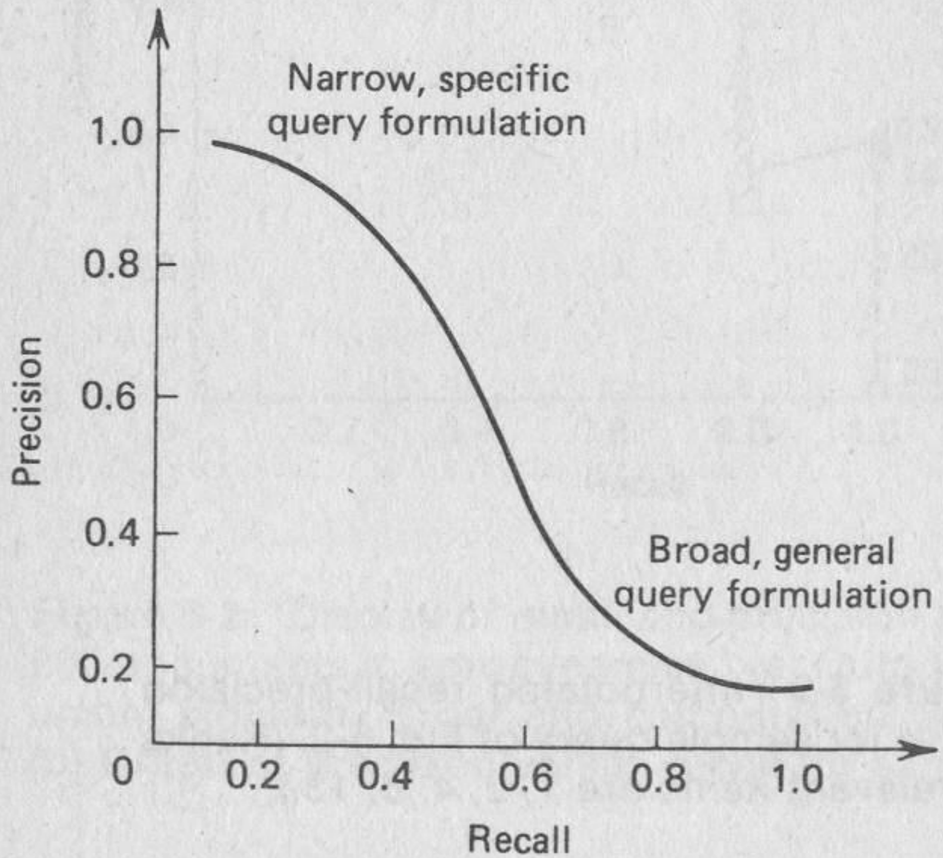
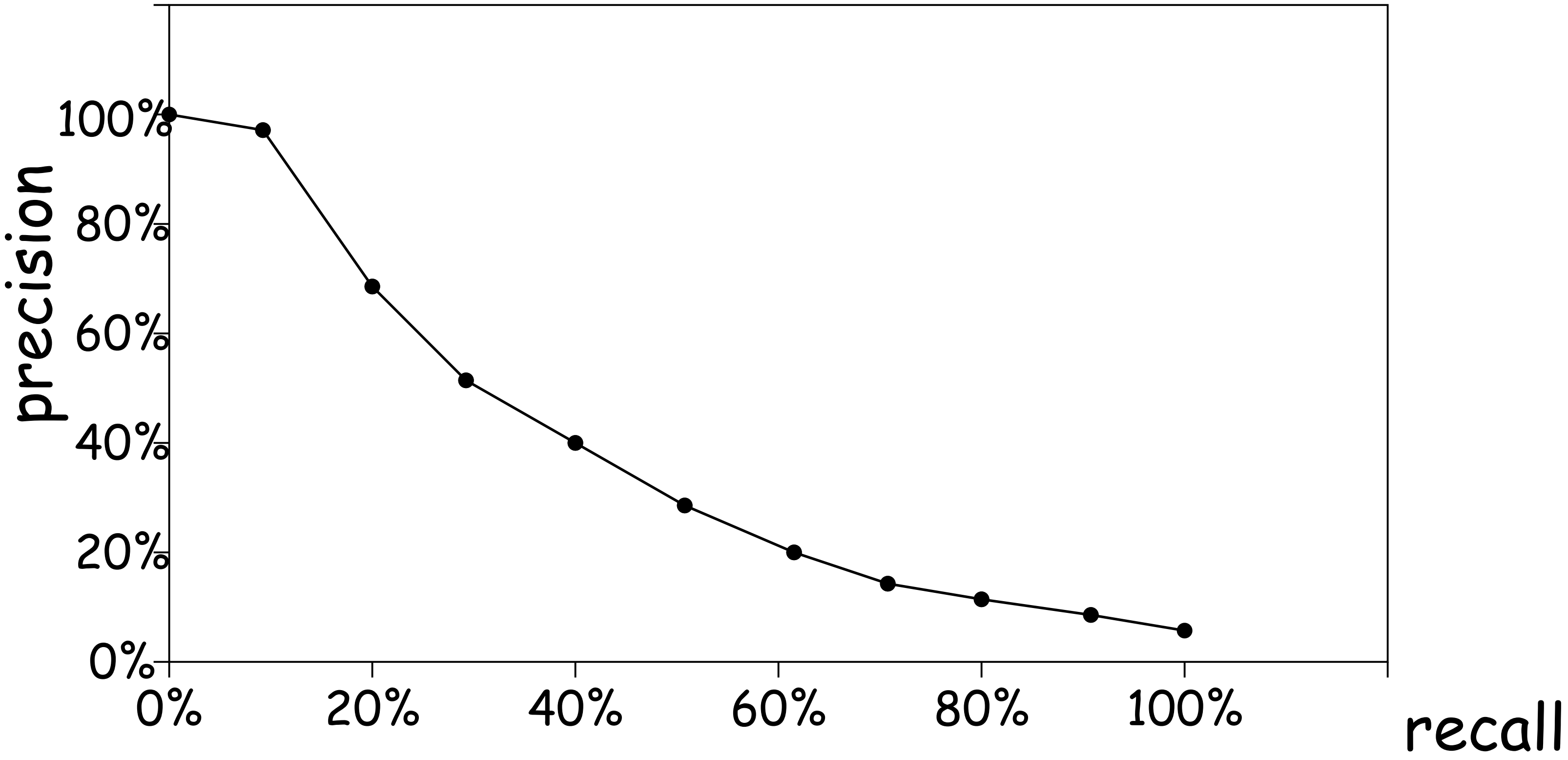


Figure 5-4 Typical average recall-precision graph.

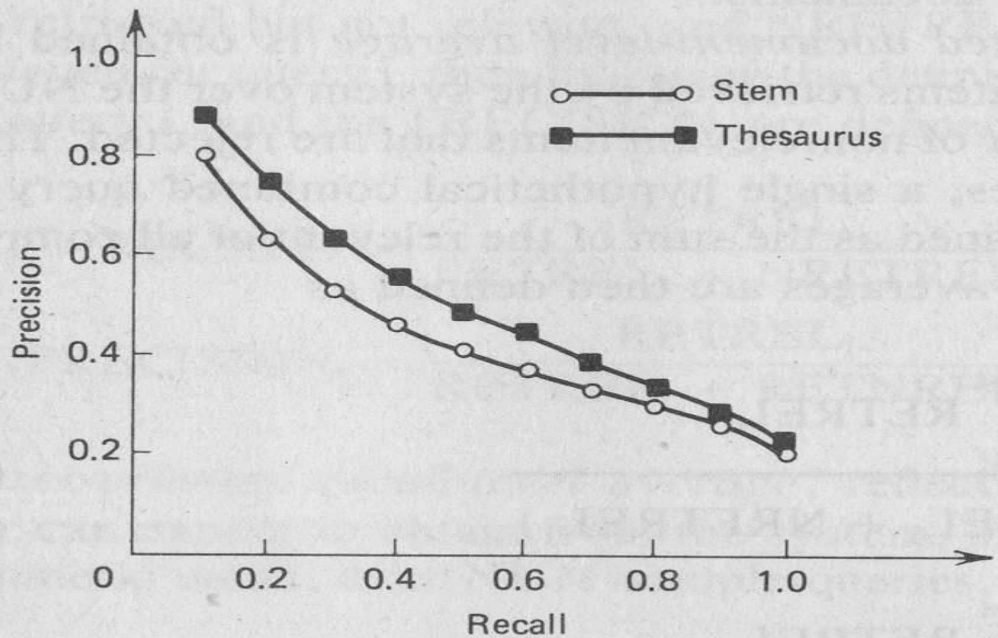
5.2 Evaluation of Retrieval: Effectiveness

P&R curve: measure precision at different levels of recall. usually, precision at 11 recall levels (0%, 10%, 20%, …, 100%)



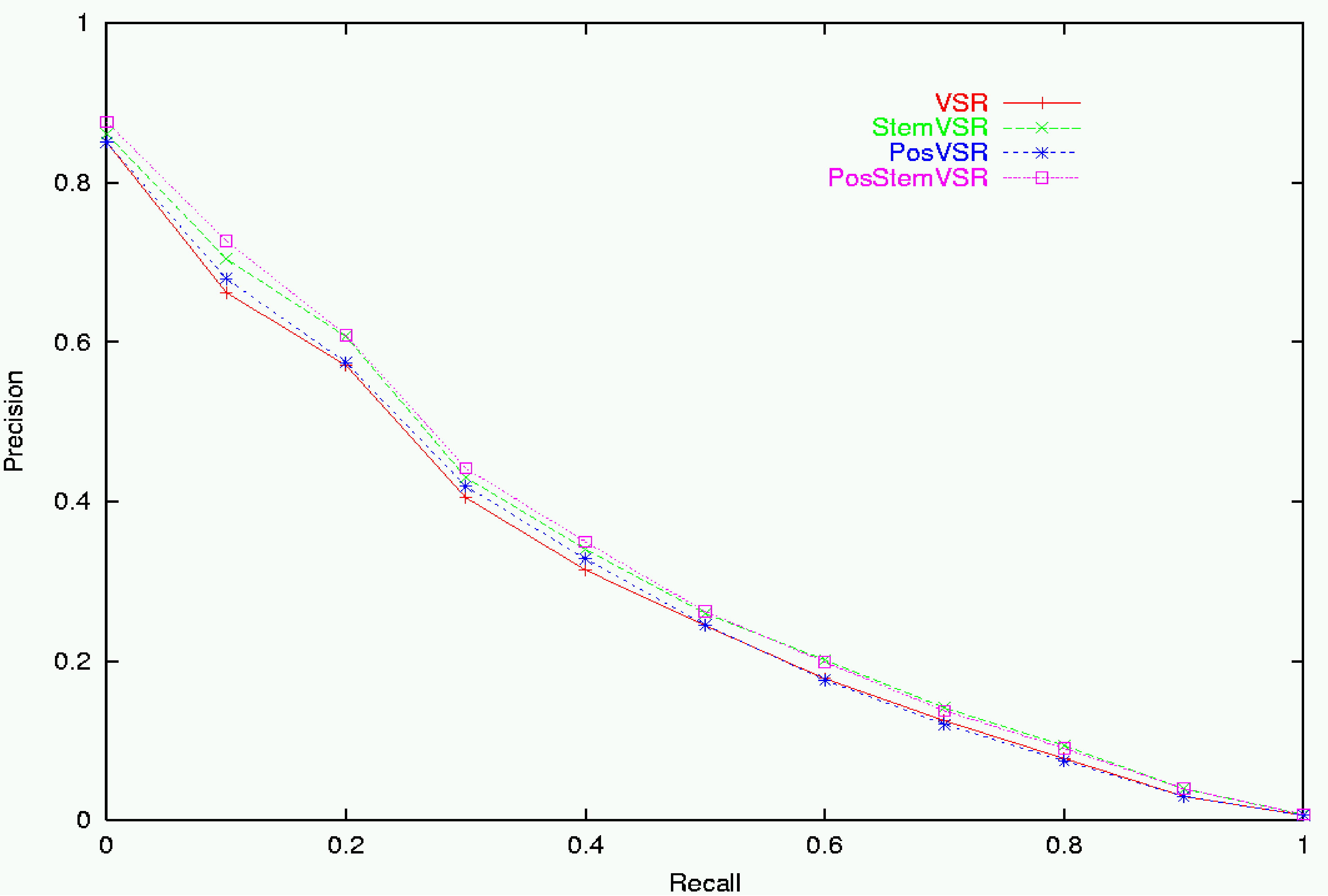
Recall	Average precision for 35 queries: Word stem	Average precision for 35 queries: Thesaurus	Improvement, %
0.1	0.7963	0.8788	10.4
0.2	0.6350	0.7567	19.2
0.3	0.5283	0.6464	22.4
0.4	0.4603	0.5577	21.2
0.5	0.4051	0.4912	21.3
0.6	0.3699	0.4470	20.8
0.7	0.3383	0.3893	15.1
0.8	0.2996	0.3287	9.7
0.9	0.2568	0.2726	6.2
1.0	0.2018	0.2093	3.7

(a)



(b)

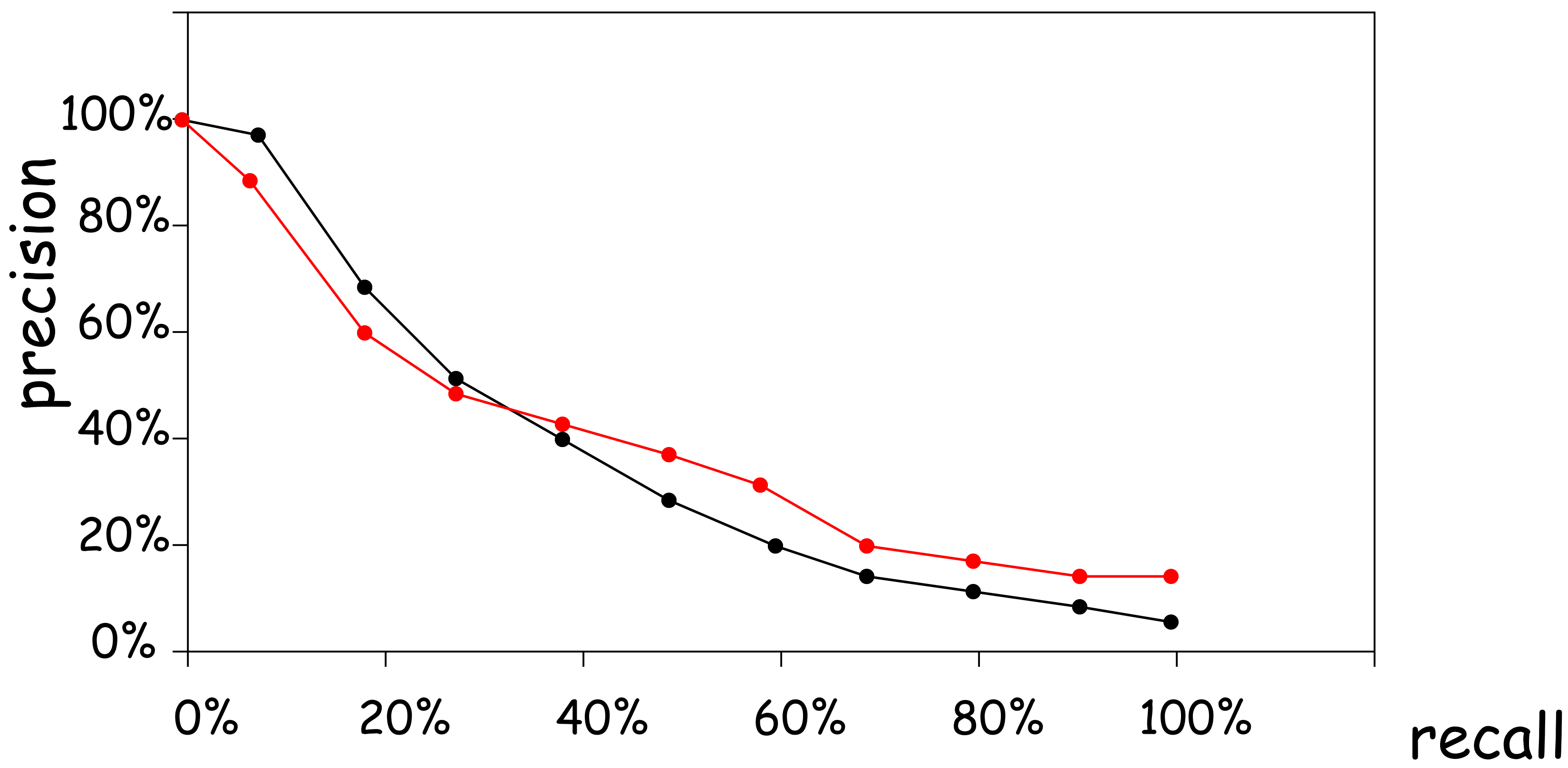
Figure 5-5 Average recall-precision results for two indexing methods (82 documents, 35 queries). (a) Recall-precision average. (b) Recall-precision graph.



Sample Recall/Precision Curve

5.2 Evaluation of Retrieval: Effectiveness

Which system performs better?



5.2 Evaluation of Retrieval: Effectiveness

F-Measure

One measure of performance that takes into account both recall and precision.

Introduced by van Rijbergen, 1979

Harmonic mean of recall and precision:

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

5.2 Evaluation of Retrieval: Effectiveness

power mean (Defi

The r-th power mean of the numbers $x_1, x_2, \ldots, x_n$ is defined as:

$$M^r(x_1, x_2, \ldots, x_n) = \left(\frac{x_1^r + x_2^r + \cdots + x_n^r}{n}\right)^{1/r}.$$

The arithmetic mean is a special case when $r = 1$. The power mean is a continuous function of r, and taking limit when $r \to 0$ gives us the geometric mean:

$$M^0(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 x_2 \cdots x_n}.$$

Also, when $r = -1$ we get

$$M^{-1}(x_1, x_2, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

the harmonic mean.

5.2 Evaluation of Retrieval: Effectiveness

harmonic mean (Definition)

If $a_1, a_2, \ldots, a_n$ are positive numbers, we define their *harmonic mean* as the inverse number of the arithmetic mean of their inverse numbers:

$$H.M. = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}}$$

- If you travel from city A to city B at x miles per hour, and then you travel back at y miles per hour. What was the average velocity for the whole trip?
  The harmonic mean of x and y!. That is, the average velocity is
  $$\frac{2}{\frac{1}{x} + \frac{1}{y}} = \frac{2xy}{x+y}.$$
- If one draws through the intersecting point of the diagonals of a trapezoid a line parallel to the parallel sides of the trapezoid, then the segment of the line inside the trapezoid is equal to the harmonic mean of the parallel sides.
- In the harmonic series
  $$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$$
  every term equals to the harmonic mean of the term preceding it and the term following it.

5.2 Evaluation of Retrieval: Effectiveness

arithmetic-geometric-harmonic means inequality (Theorem)

Let $x_1, x_2, \ldots, x_n$ be positive numbers. Then

$$\begin{aligned}\max\{x_1, x_2, \ldots, x_n\} &\geq \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &\geq \sqrt[n]{x_1 x_2 \cdots x_n} \\ &\geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \\ &\geq \min\{x_1, x_2, \ldots, x_n\}\end{aligned}$$

The equality is obtained if and only if $x_1 = x_2 = \cdots = x_n$.

5.2 Evaluation of Retrieval: Effectiveness

general means inequality (Theorem)

The power means inequality is a generalization of arithmetic-geometric means inequality.

If $0 \neq r \in \mathbb{R}$, the r-mean (or r-th power mean) of the nonnegative numbers $a_1, \ldots, a_n$ is defined as

$$M^r(a_1, a_2, \ldots, a_n) = \left(\frac{1}{n} \sum_{k=1}^{n} a_k^r \right)^{1/r}$$

Given real numbers x, y such that $xy \neq 0$ and $x < y$, we have

$$M^x \leq M^y$$

and the equality holds if and only if $a_1 = \ldots = a_n$.

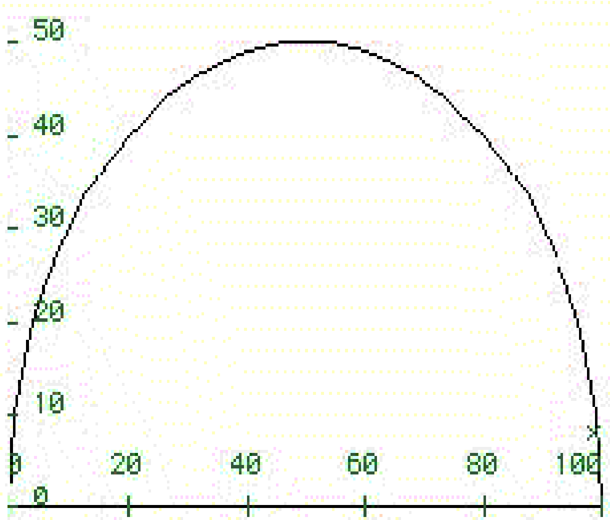
Additionally, if we define M^0 to be the geometric mean $(a_1 a_2 \ldots a_n)^{1/n}$, we have that the inequality above holds for arbitrary real numbers $x < y$.

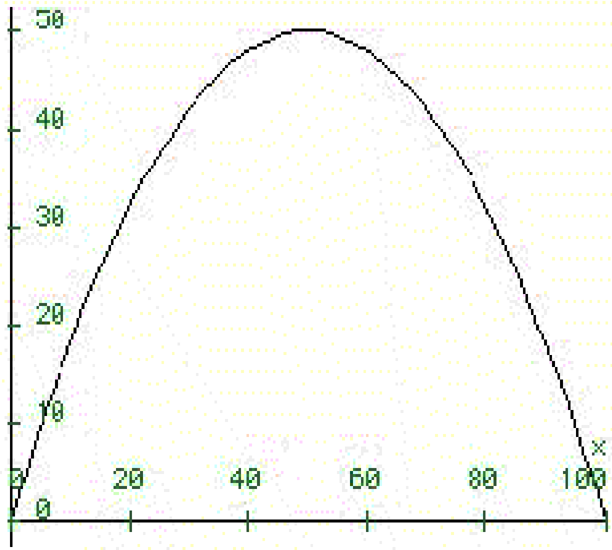
The mentioned inequality is a special case of this one, since M^1 is the arithmetic mean, M^0 is the geometric mean and M^{-1} is the harmonic mean.

This inequality can be further generalized using weighted power means.

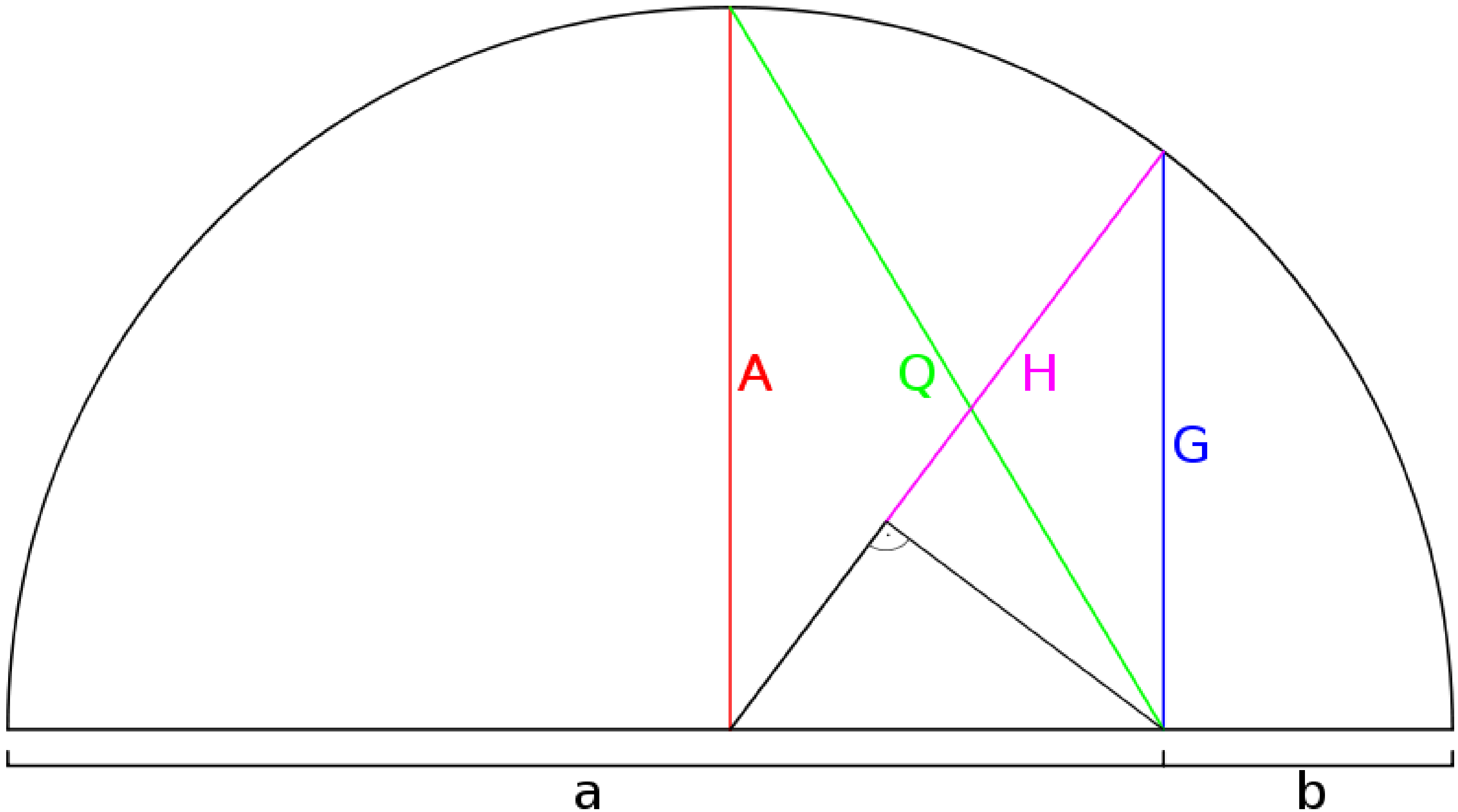
5.2 Evaluation of Retrieval: Effectiveness

x	y	arithmetic mean	geometric mean	harmonic mean
50	50	50	50	50
40	60	50	49	48
30	70	50	46	42
20	80	50	40	32





5.2 Evaluation of Retrieval: Effectiveness



Geometrical representation of common mathematical means. a,b-two scalars. A=Arithmetic mean of scalars 'a' and 'b'. G=Geometric mean, H=Harmonic mean, Q=Quadratic mean (Root mean square)

5.2 Evaluation of Retrieval: Effectiveness

- cannot take mean of P&R
  - if R = 50% P = 50% M = 50%
  - if R = 100% P = 10% M = 55% (not fair)
- take harmonic mean

$$HM = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

HM is high only when both P&R are high

if R = 50% and P = 50% HM = 50%

if R = 100% and P = 10% HM = 18.2%

5.2 Evaluation of Retrieval: Effectiveness

weighted power mean (Definition)

If $w_1, w_2, \ldots, w_n$ are positive real numbers such that $w_1 + w_2 + \cdots + w_n = 1$, we define the r-*th weighted power mean* of the x_i as:

$$M_w^r(x_1, x_2, \ldots, x_n) = \left(w_1 x_1^r + w_2 x_2^r + \cdots + w_n x_n^r\right)^{1/r}.$$

When all the $w_i = \frac{1}{n}$ we get the standard power mean. The weighted power mean is a continuous function of r, and taking limit when $r \to 0$ gives us

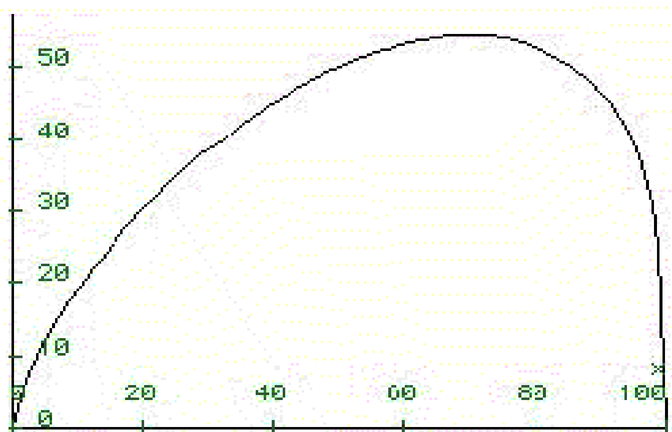
$$M_w^0 = x_1^{w_1} x_2^{w_2} \cdots w_n^{w_n}.$$

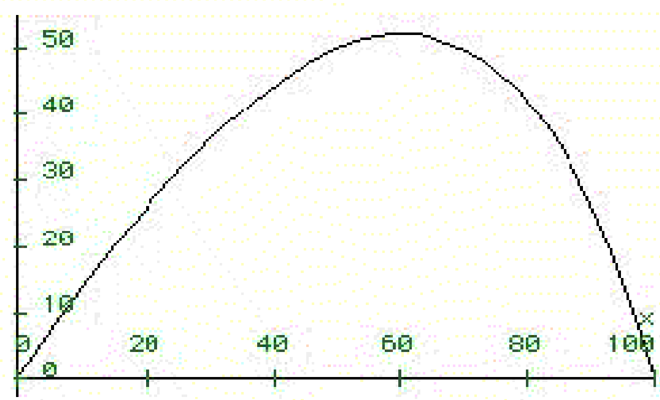
We can weighted use power means to generalize the power means inequality. If w is a set of weights, and if $r < s$ then

$$M_w^r \leq M_w^s.$$

Mean	*Formula*
weighted arithmetic mean of x and y	$0.7x + 0.3y$
weighted geometric mean of x and y	$x^{0.7} \times y^{0.3}$
weighted harmonic mean of x and y	$1/(0.7/x + 0.3/y) = xy/(0.7y + 0.3x)$

x	y	weighted arithmetic mean	weighted geometric mean	weighted harmonic mean
80	20	62	53	42
70	30	58	54	50
60	40	54	53	52
50	50	50	50	50
40	60	46	45	44
30	70	42	37	36
20	80	38	30	26





A combined measure: *F*

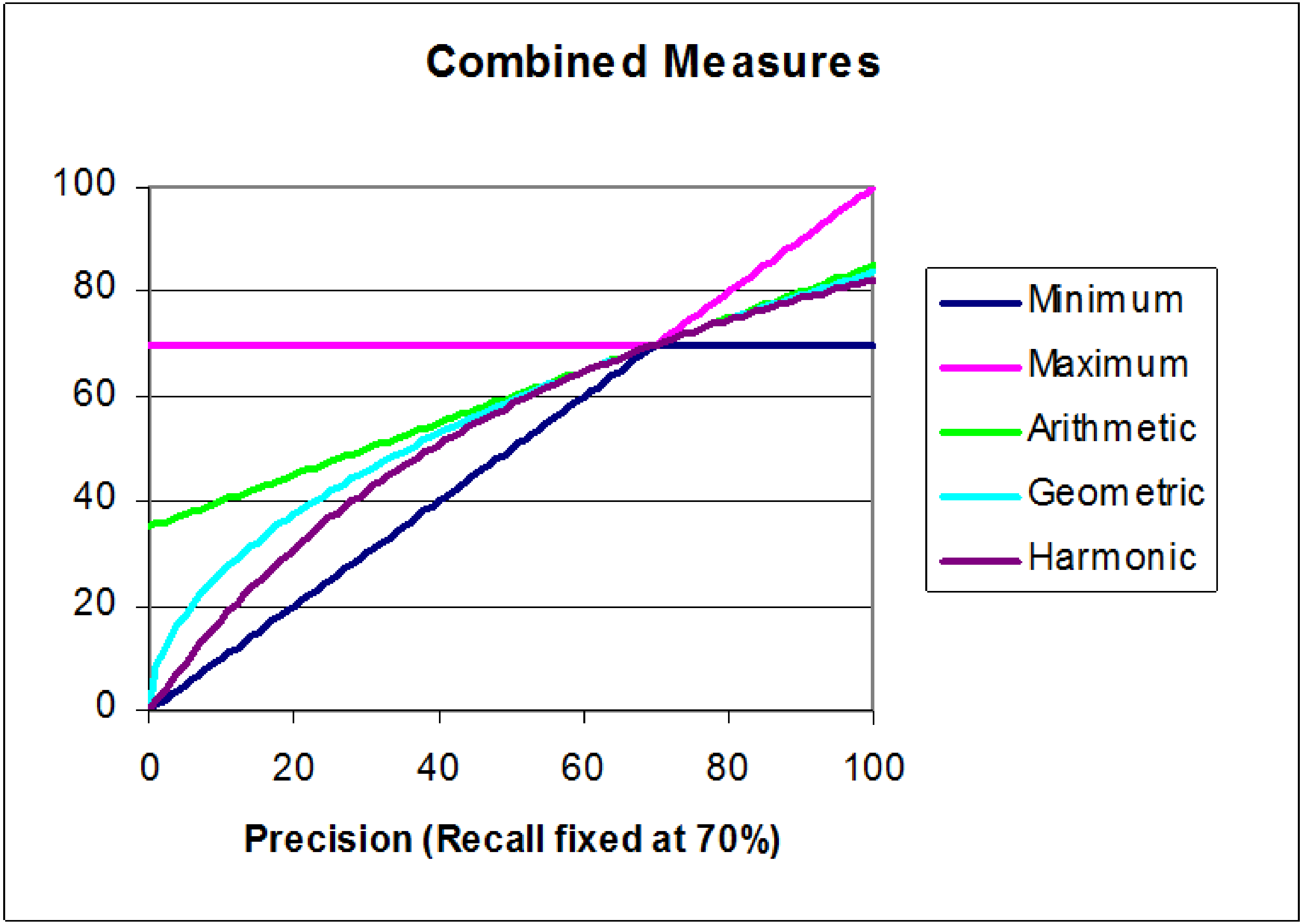
- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

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

sqr(β) = （1− *α*）/*α*

- People usually use balanced F_1 measure
  - i.e., with β = 1 or α = ½
- Harmonic mean is a conservative average

F_1 and other averages



5.2 Evaluation of Retrieval: Effectiveness

Other evaluations:

-- Break-even point R=P

-- Document cutoff levels

Web search: R? and

P&R do not evaluate the ranking

d_{123} √ d_{84} × ≡ d_{84} × d_{123} √

...

5.2 Evaluation of Retrieval: Effectiveness

R-precision (the precision at the R-th position in the ranking)

Fix the number of documents retrieved at several levels

ex. top 5, top 10, top 20, top 100, top 500…

Measure precision at each of these levels

	system 1	*system 2*	*system 3*
	d1 √	d10 ×	d6 ×
	d2 √	d9 ×	d1 √
	d3 √	d8 ×	d2 √
	d4 √	d7 ×	d10 ×
	d5 √	d6 ×	d9 ×
	d6 ×	d1 √	d3 √
	d7 ×	d2 √	d5 √
	d8 ×	d3 √	d4 √
	d9 ×	d4 √	d7 ×
	d10 ×	d5 √	d8 ×
precision at 5	1.0	0.0	0.4
precision at 10	0.5	0.5	0.5

Kappa系数：衡量判断（标注）的一致性

A \ B	Yes	No
Yes	a	b
No	c	d

A \ B	Yes	No
Yes	20	5
No	10	15

Cohen's kappa coefficient (*κ*)

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

*Kappa*系数：衡量判断（标注）的一致性

相关概率知识：

一个事件的概率：$p(x)$

两个事件的概率（相互独立条件下）$p(xy)=p(x)*p(y)$

两个事件的条件概率：$p(y|x)$

两个事件的联合概率： $p(xy)=p(x)*p(y|x)$

B / A	Yes	No
Yes	a	b
No	c	d

B / A	Yes	No
Yes	20	5
No	10	15

The observed proportionate agreement is:

$$p_o = \frac{a+d}{a+b+c+d} = \frac{20+15}{50} = 0.7$$

To calculate p_e (the probability of random agreement) we note that:

- Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
- Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

So the expected probability that both would say yes at random is:

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$$

Similarly:

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$$

Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

*Kappa*系数：衡量判断（标注）的一致性

- Kappa > 0.8 = good agreement
- Depends on purpose of study
- 0.67 < Kappa < 0.8 -> “tentative conclusions” (Carletta 96)

- For > 2 judges: average pairwise Kappas

5.3 Evaluation of Retrieval: Efficiency

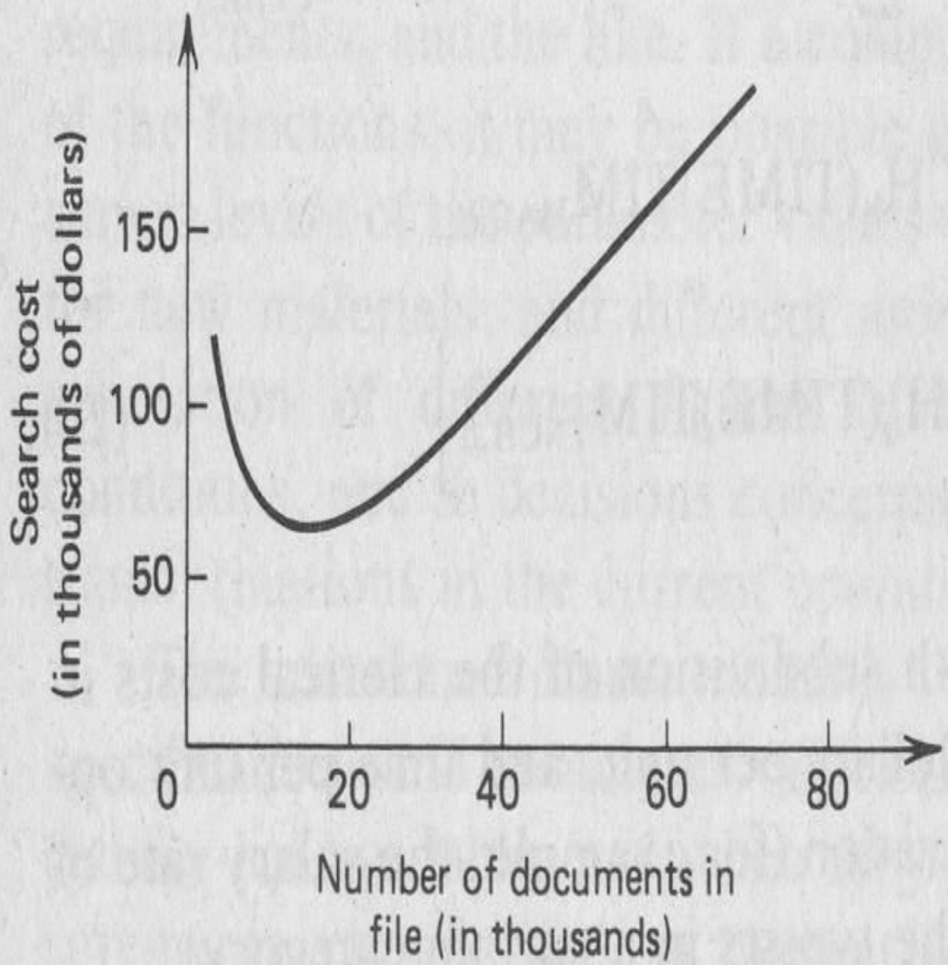


Figure 5-12 Typical cost curve reflecting search cost. (*Adapted from reference 77.*)