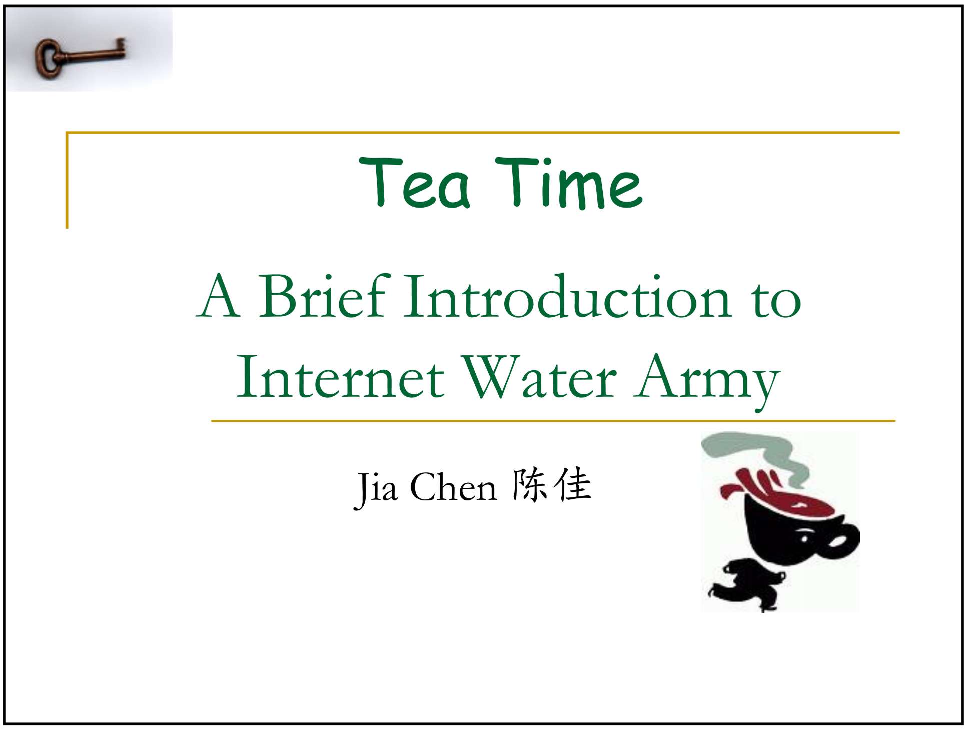
Welcome to the class of Advanced Topics in Information Retrieval !

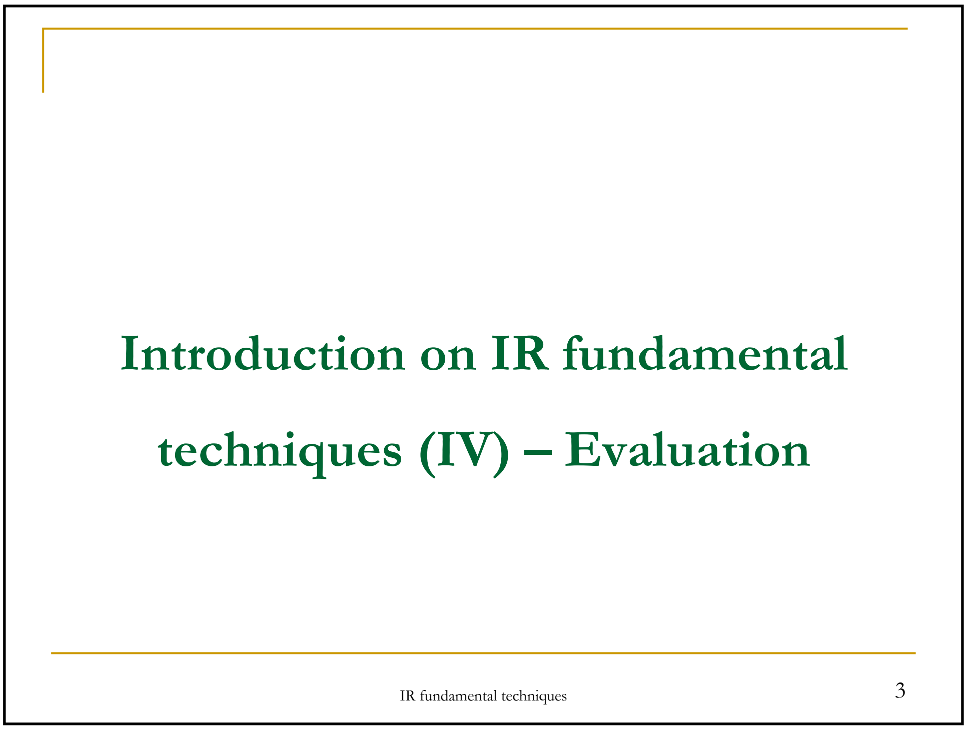
Min ZHANG （张敏）
z-m@tsinghua.edu.cn



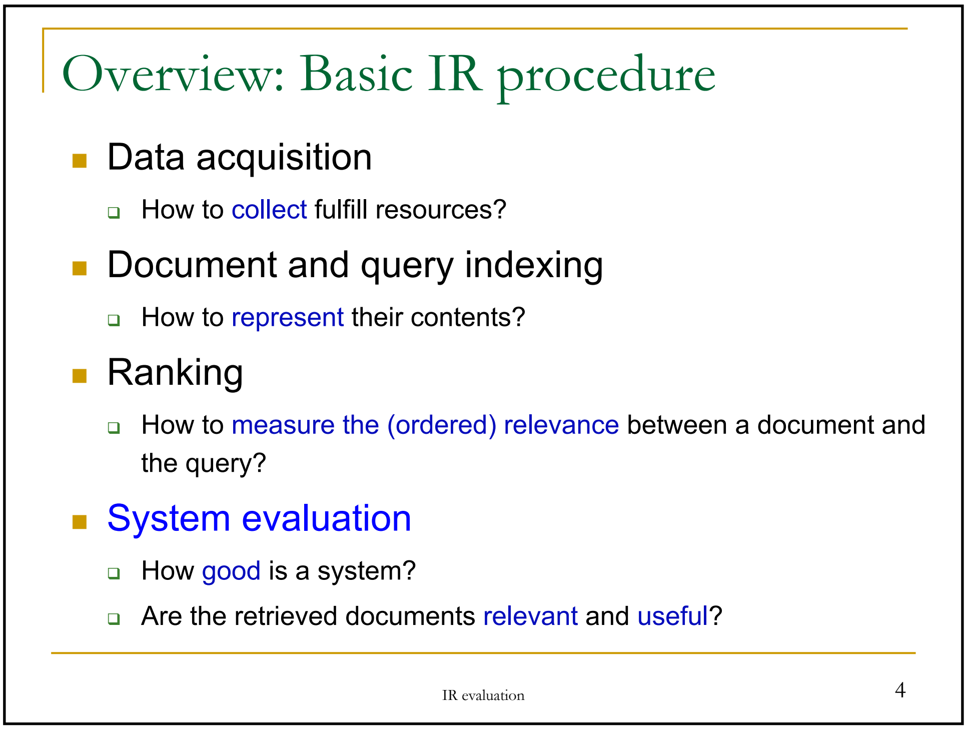
Tea Time

A Brief Introduction to Internet Water Army

Jia Chen 陈佳



Introduction on IR fundamental techniques (IV) – Evaluation



Overview: Basic IR procedure

- Data acquisition
  - How to collect fulfill resources?
- Document and query indexing
  - How to represent their contents?
- Ranking
  - How to measure the (ordered) relevance between a document and the query?
- System evaluation
  - How good is a system?
  - Are the retrieved documents relevant and useful?

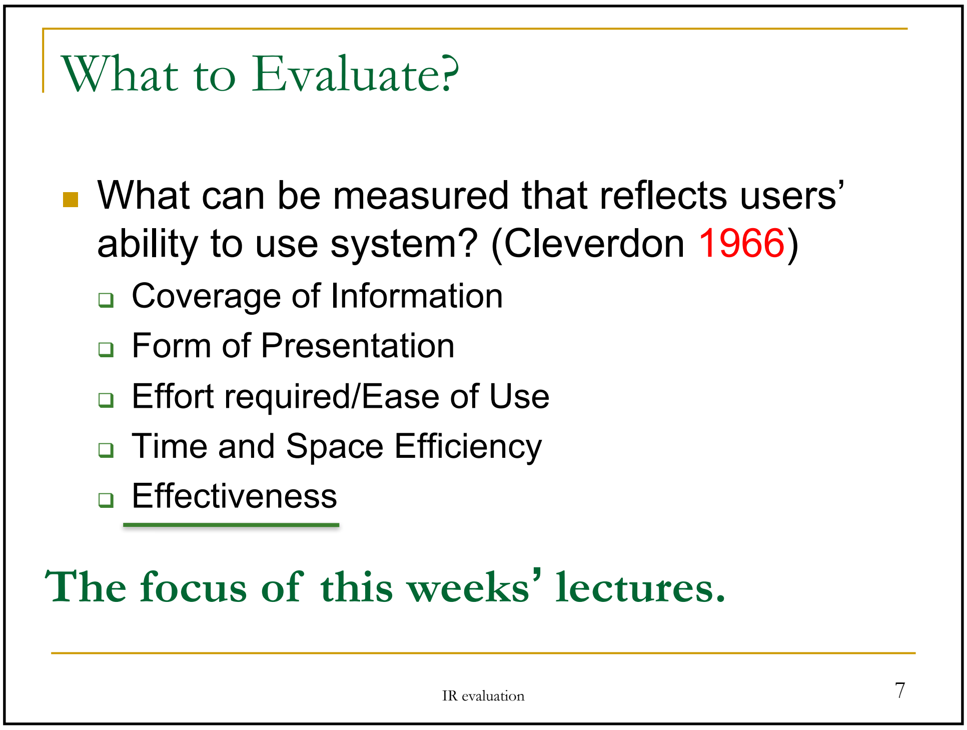
"Evaluation is a major force in research, development and applications related to information retrieval."

"Evaluation became central to R&D to such an extent that new designs and proposals and their evaluation became one."

-- Saracevic, T. (1995). Evaluation of evaluation in information retrieval. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Special issue of SIGIR Forum, 138-146

Evaluation

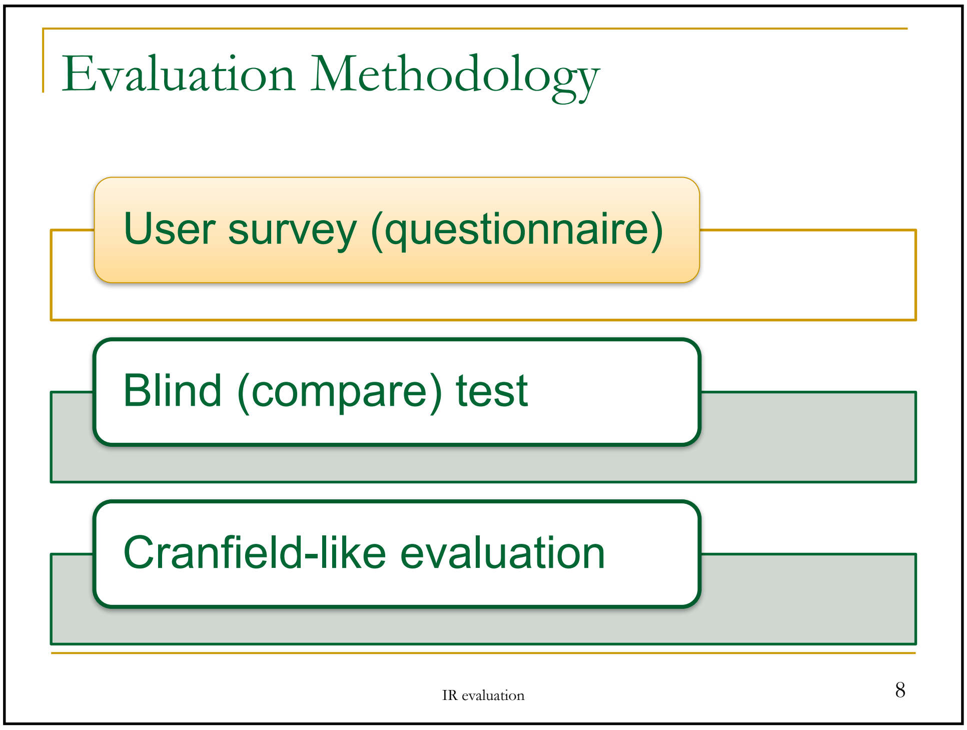
- Evaluation methodology
- Standard benchmarks for IR
- Annotation Consistency
- Common metrics for relevance



What to Evaluate?

- What can be measured that reflects users' ability to use system? (Cleverdon 1966)
  - Coverage of Information
  - Form of Presentation
  - Effort required/Ease of Use
  - Time and Space Efficiency
  - Effectiveness

The focus of this weeks' lectures.



Evaluation Methodology

User survey (questionnaire)

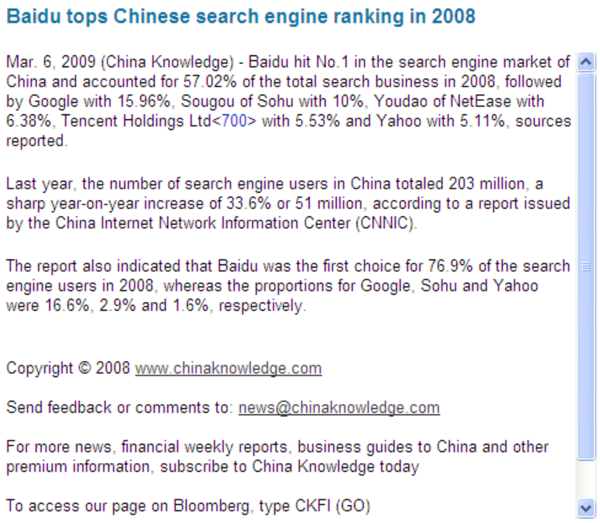
Blind (compare) test

Cranfield-like evaluation

Evaluation methodology – user survey

(user statistics)

Search Engine News Review (Mar. 04 – Mar. 11, 2009)



Baidu tops Chinese search engine ranking in 2008

Mar. 6, 2009 (China Knowledge) - Baidu hit No.1 in the search engine market of China and accounted for 57.02% of the total search business in 2008, followed by Google with 15.96%, Sougou of Sohu with 10%, Youdao of NetEase with 6.38%, Tencent Holdings Ltd<700> with 5.53% and Yahoo with 5.11%, sources reported.

Last year, the number of search engine users in China totaled 203 million, a sharp year-on-year increase of 33.6% or 51 million, according to a report issued by the China Internet Network Information Center (CNNIC).

The report also indicated that Baidu was the first choice for 76.9% of the search engine users in 2008, whereas the proportions for Google, Sohu and Yahoo were 16.6%, 2.9% and 1.6%, respectively.

Copyright © 2008 www.chinaknowledge.com

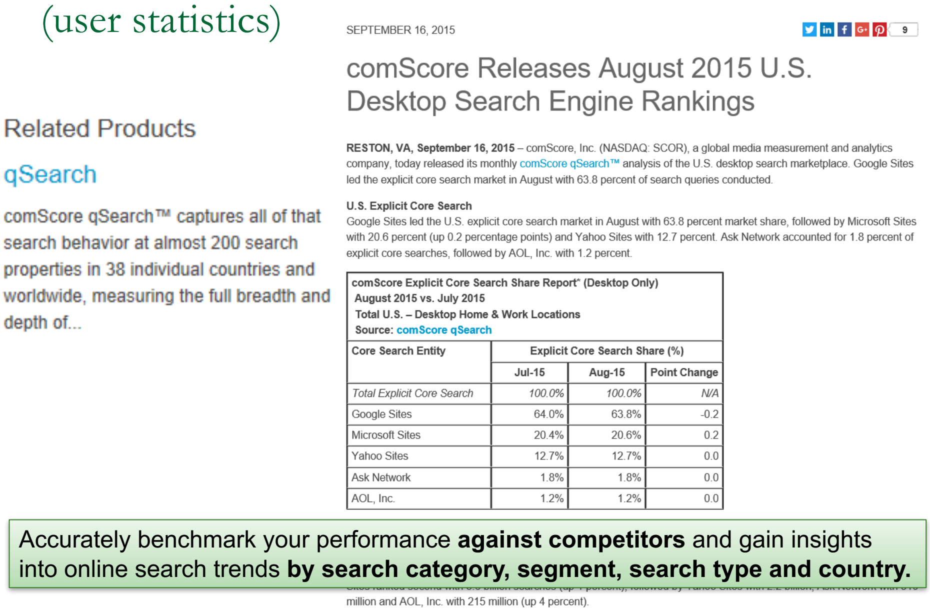
Send feedback or comments to: news@chinaknowledge.com

For more news, financial weekly reports, business guides to China and other premium information, subscribe to China Knowledge today

To access our page on Bloomberg, type CKFI (GO)

Evaluation methodology – user survey

(user statistics)



Related Products

qSearch

comScore qSearch™ captures all of that search behavior at almost 200 search properties in 38 individual countries and worldwide, measuring the full breadth and depth of...

SEPTEMBER 16, 2015

comScore Releases August 2015 U.S. Desktop Search Engine Rankings

RESTON, VA, September 16, 2015 – comScore, Inc. (NASDAQ: SCOR), a global media measurement and analytics company, today released its monthly comScore qSearch™ analysis of the U.S. desktop search marketplace. Google Sites led the explicit core search market in August with 63.8 percent of search queries conducted.

U.S. Explicit Core Search

Google Sites led the U.S. explicit core search market in August with 63.8 percent market share, followed by Microsoft Sites with 20.6 percent (up 0.2 percentage points) and Yahoo Sites with 12.7 percent. Ask Network accounted for 1.8 percent of explicit core searches, followed by AOL, Inc. with 1.2 percent.

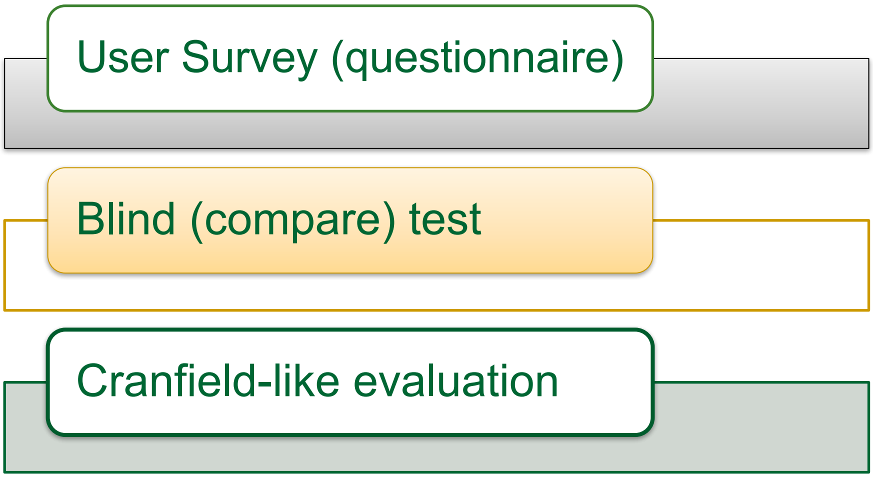
comScore Explicit Core Search Share Report* (Desktop Only)
August 2015 vs. July 2015
Total U.S. – Desktop Home & Work Locations
Source: comScore qSearch

Core Search Entity	Explicit Core Search Share (%)		
	Jul-15	Aug-15	Point Change
Total Explicit Core Search	*100.0%*	*100.0%*	*N/A*
Google Sites	64.0%	63.8%	-0.2
Microsoft Sites	20.4%	20.6%	0.2
Yahoo Sites	12.7%	12.7%	0.0
Ask Network	1.8%	1.8%	0.0
AOL, Inc.	1.2%	1.2%	0.0

Accurately benchmark your performance **against competitors** and gain insights into online search trends **by search category, segment, search type and country.**

million and AOL, Inc. with 215 million (up 4 percent).

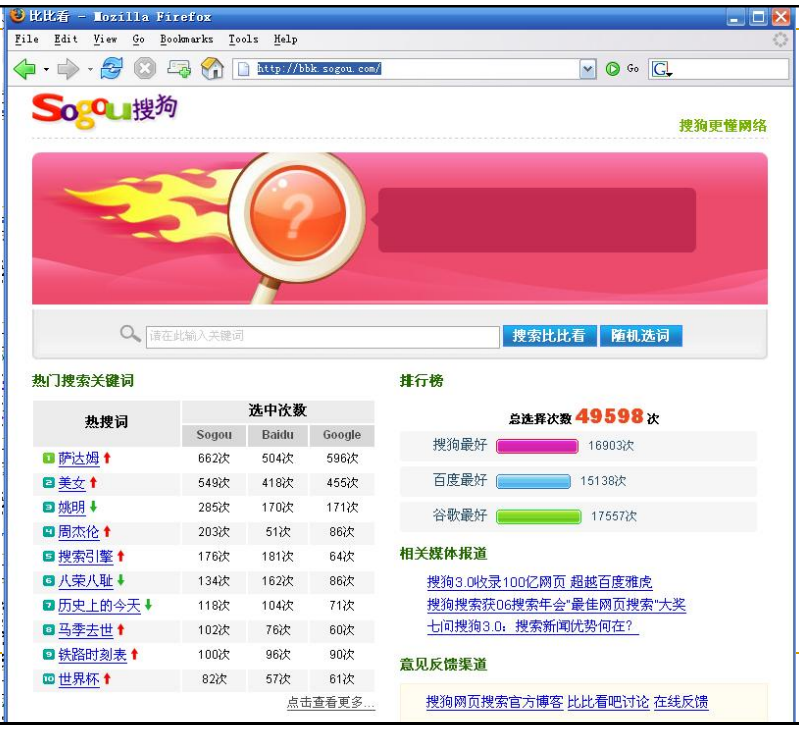
Evaluation Methodology



User Survey (questionnaire)

Blind (compare) test

Cranfield-like evaluation



热门搜索关键词

热搜词	选中次数		
	Sogou	Baidu	Google
1 萨达姆 ↑	662次	504次	596次
2 美女 ↑	549次	418次	455次
3 姚明 ↓	285次	170次	171次
4 周杰伦 ↑	203次	51次	86次
5 搜索引擎 ↑	176次	181次	64次
6 八荣八耻 ↓	134次	162次	86次
7 历史上的今天 ↓	118次	104次	71次
8 马季去世 ↑	102次	76次	60次
9 铁路时刻表 ↑	100次	96次	90次
10 世界杯 ↑	82次	57次	61次

点击查看更多...

排行榜

总选择次数 49598 次

搜狗最好	16903次
百度最好	15138次
谷歌最好	17557次

相关媒体报道

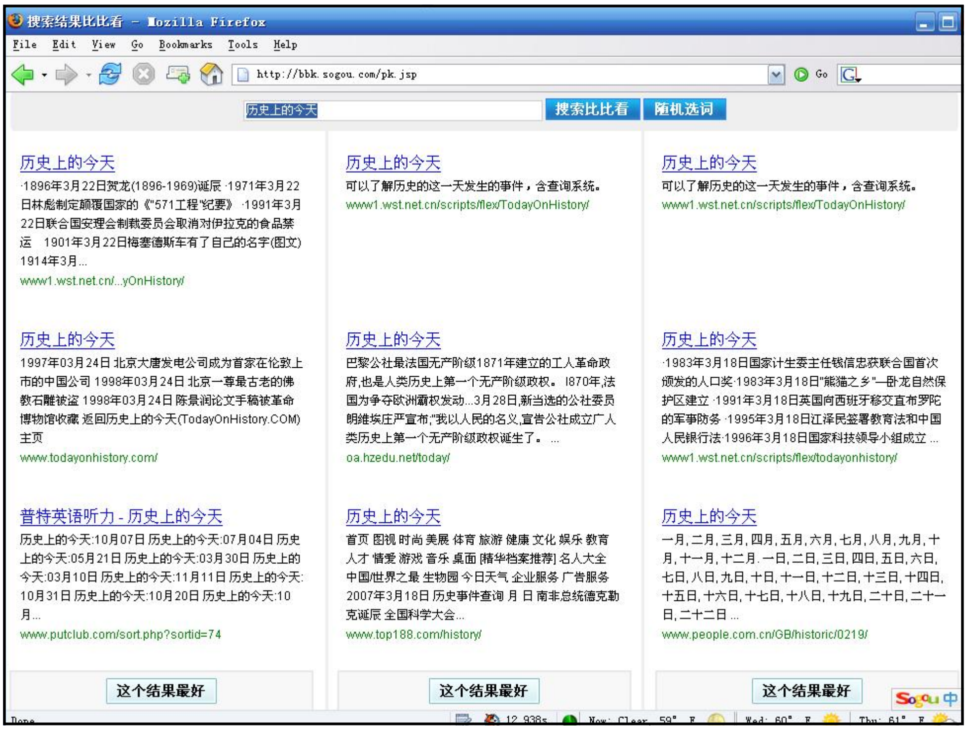
搜狗3.0收录100亿网页 超越百度雅虎

搜狗搜索获06搜索年会"最佳网页搜索"大奖

七问搜狗3.0：搜索新闻优势何在？

意见反馈渠道

搜狗网页搜索官方博客 比比看吧讨论 在线反馈


搜索结果比比看 - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://bbk.sogou.com/pk.jsp
Go
历史上的今天
搜索比比看
随机选词
Done
12.938s
Now: Clear, 59° F
Wed: 60° F
Thu: 61° F
Sogou 中

历史上的今天
·1896年3月22日贺龙(1896-1969)诞辰 ·1971年3月22日林彪制定颠覆国家的《"571工程"纪要》 ·1991年3月22日联合国安理会制裁委员会取消对伊拉克的食品禁运 1901年3月22日梅塞德斯车有了自己的名字(图文) 1914年3月...
www1.wst.net.cn/...yOnHistory/

历史上的今天
1997年03月24日 北京大唐发电公司成为首家在伦敦上市的中国公司 1998年03月24日 北京一尊最古老的佛教石雕被盗 1998年03月24日 陈景润论文手稿被革命博物馆收藏 返回历史上的今天(TodayOnHistory.COM)主页
www.todayonhistory.com/

普特英语听力 - 历史上的今天
历史上的今天:10月07日 历史上的今天:07月04日 历史上的今天:05月21日 历史上的今天:03月30日 历史上的今天:03月10日 历史上的今天:11月11日 历史上的今天:10月31日 历史上的今天:10月20日 历史上的今天:10月...
www.putclub.com/sort.php?sortid=74

这个结果最好

历史上的今天
可以了解历史的这一天发生的事件，含查询系统。
www1.wst.net.cn/scripts/flex/TodayOnHistory/

历史上的今天
巴黎公社最法国无产阶级1871年建立的工人革命政府,也是人类历史上第一个无产阶级政权。 I870年,法国为争夺欧洲霸权发动...3月28日,新当选的公社委员朗維埃庄严宣布;"我以人民的名义,宣告公社成立广人类历史上第一个无产阶级政权诞生了。 ...
oa.hzedu.net/today/

历史上的今天
首页 图视 时尚 美展 体育 旅游 健康 文化 娱乐 教育 人才 情爱 游戏 音乐 桌面 [精华档案推荐] 名人大全 中国/世界之最 生物园 今日天气 企业服务 广告服务 2007年3月18日 历史事件查询 月 日 南非总统德克勒克诞辰 全国科学大会...
www.top188.com/history/

这个结果最好

历史上的今天
可以了解历史的这一天发生的事件，含查询系统。
www1.wst.net.cn/scripts/flex/TodayOnHistory/

历史上的今天
·1983年3月18日国家计生委主任钱信忠获联合国首次颁发的人口奖·1983年3月18日"熊猫之乡"—卧龙自然保护区建立 ·1991年3月18日英国向西班牙移交直布罗陀的军事防务 ·1995年3月18日江泽民签署教育法和中国人民银行法·1996年3月18日国家科技领导小組成立 ...
www1.wst.net.cn/scripts/flex/todayonhistory/

历史上的今天
一月, 二月, 三月, 四月, 五月, 六月, 七月, 八月, 九月, 十月, 十一月, 十二月. 一日, 二日, 三日, 四日, 五日, 六日, 七日, 八日, 九日, 十日, 十一日, 十二日, 十三日, 十四日, 十五日, 十六日, 十七日, 十八日, 十九日, 二十日, 二十一日, 二十二日 ...
www.people.com.cn/GB/historic/0219/

这个结果最好

Blind test is not always reliable

Microsoft Faces Branding Problem In Effort to Top Google

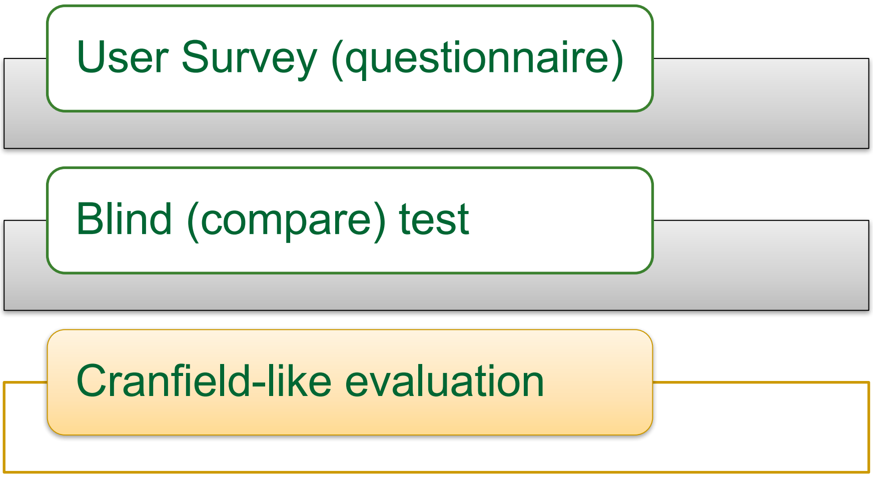
http://blogs.wsj.com/digits/2009/04/08/microsoft-faces-branding-problem-in-effort-to-top-google/ Wall Street Journal, April 8, 2009

Yusuf Mehdi

During regular "blind taste tests," in which Microsoft asks randomly-selected consumers to score the quality of results from various Internet search engines, the quality of Microsoft's search results have so improved that people can't tell the difference between Microsoft and Google search results, says Mr. Mehdi, senior vice president of Microsoft's online audience business group. But when Microsoft slaps the Google brand name on the results from Microsoft's own search engine during another portion of its tests, users invariably score them highest.

" Just by putting the name up, people think it's more relevant," he says.

Evaluation Methodology



User Survey (questionnaire)

Blind (compare) test

Cranfield-like evaluation

Cranfield-like evaluation

- Proposed by Cleverdon at Cranfield,UK in 1950’s
- The 3 components of the IR test collections
  - 1) a collection of documents,
  - 2) a set of user requests or queries
  - 3) a set of relevance judgments
    - i.e. a set of documents judged to be relevant to each query
- Comparative testing

Relevance

Relevance is a context-, task-dependent property of documents

> "Relevance is the correspondence **in context** between **an information requirement** statement ... and an article (a **document**), that is, the extent to which the article covers the material that is appropriate to the requirement statement."
>
> F. W. Lancaster, 1979

Difficulties in Evaluating IR Systems

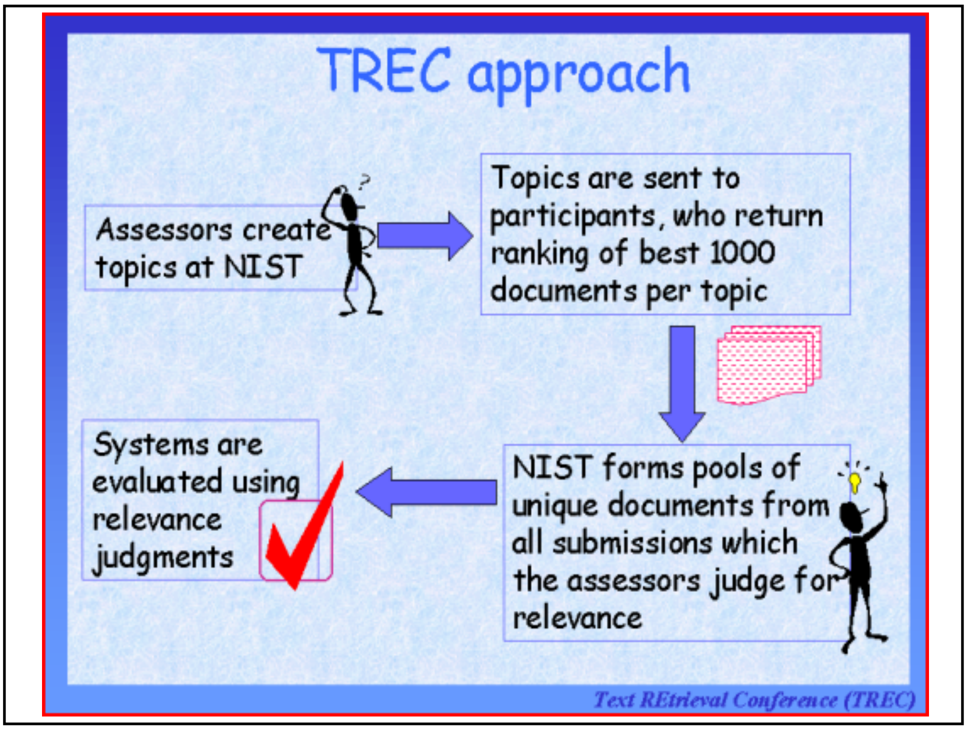
- Effectiveness is related to the ***relevancy*** of retrieved items.
- Relevancy is not typically binary but continuous.
- Even if relevancy is binary, it can be a difficult judgment to make.
- Relevancy, from a human standpoint, is:
  - *Subjective*: Depends upon a specific user's judgment.
  - *Situational*: Relates to user's current needs.
  - *Cognitive*: Depends on human perception and behavior.
  - *Dynamic*: Changes over time.

Evaluation

- Evaluation methodology
- Standard benchmarks for IR
- Annotation Consistency
- Common metrics for relevance

Standard relevance benchmarks – TREC

- Text REtrieval Conference
  - One of the most important IR benchmarks
- Sponsers: NIST and DARPA
  - National Institute of Standards and Technology
  - Defense Advanced Research Projects Agency
- 1992 (TREC1) ~ 2017（TREC26）～ ……
- http://trec.nist.gov



TREC approach

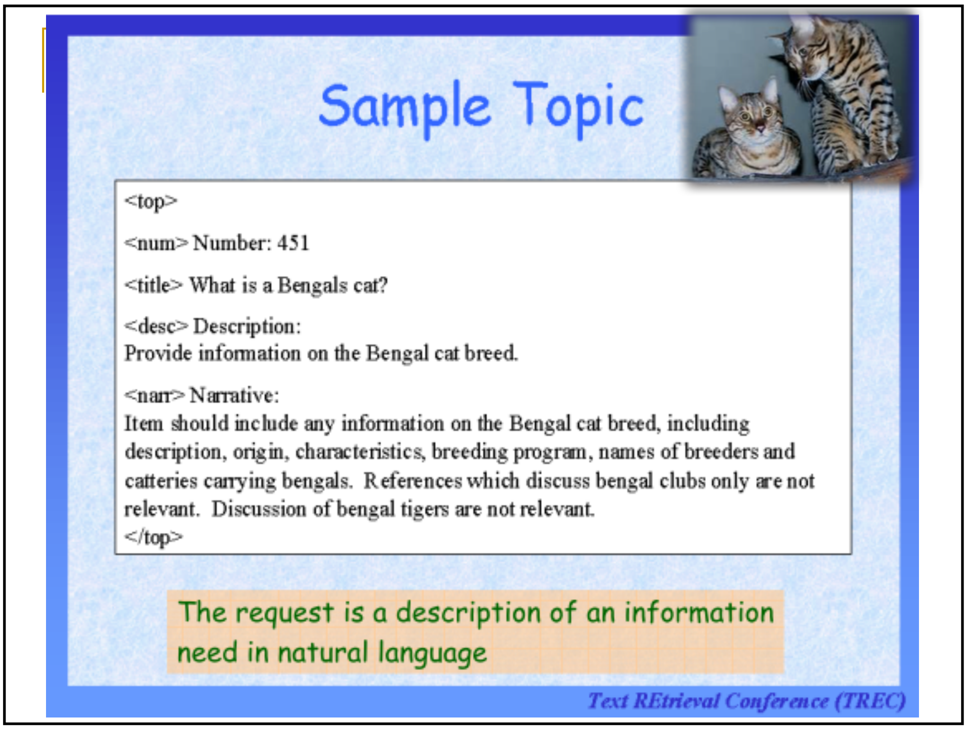
Assessors create topics at NIST

Topics are sent to participants, who return ranking of best 1000 documents per topic

NIST forms pools of unique documents from all submissions which the assessors judge for relevance

Systems are evaluated using relevance judgments

Text REtrieval Conference (TREC)



Sample Topic

```
<top>

<num> Number: 451

<title> What is a Bengals cat?

<desc> Description:
Provide information on the Bengal cat breed.

<narr> Narrative:
Item should include any information on the Bengal cat breed, including
description, origin, characteristics, breeding program, names of breeders and
catteries carrying bengals.  References which discuss bengal clubs only are not
relevant.  Discussion of bengal tigers are not relevant.
</top>
```

The request is a description of an information need in natural language

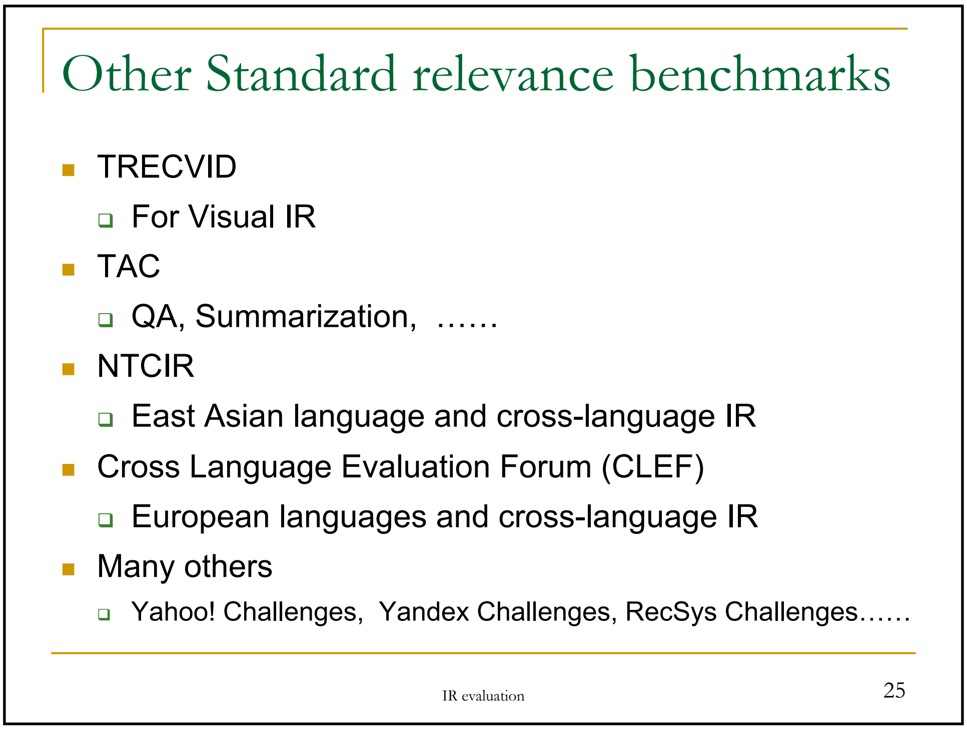
Text REtrieval Conference (TREC)

The problem of the human judgment

- Time-consuming & heavy human effort cost
  - **9 people months** are required to judge **one topic** for a collection of **8 million documents**. (Voorhees, 2001)
- TREC solution: the so-called "*pooling*" method
  - Two assumptions
    - Vast majority of relevant docs is collected in the assembled pool
    - Docs not in the pool were considered to be irrelevant
  - ***Comparative testing***

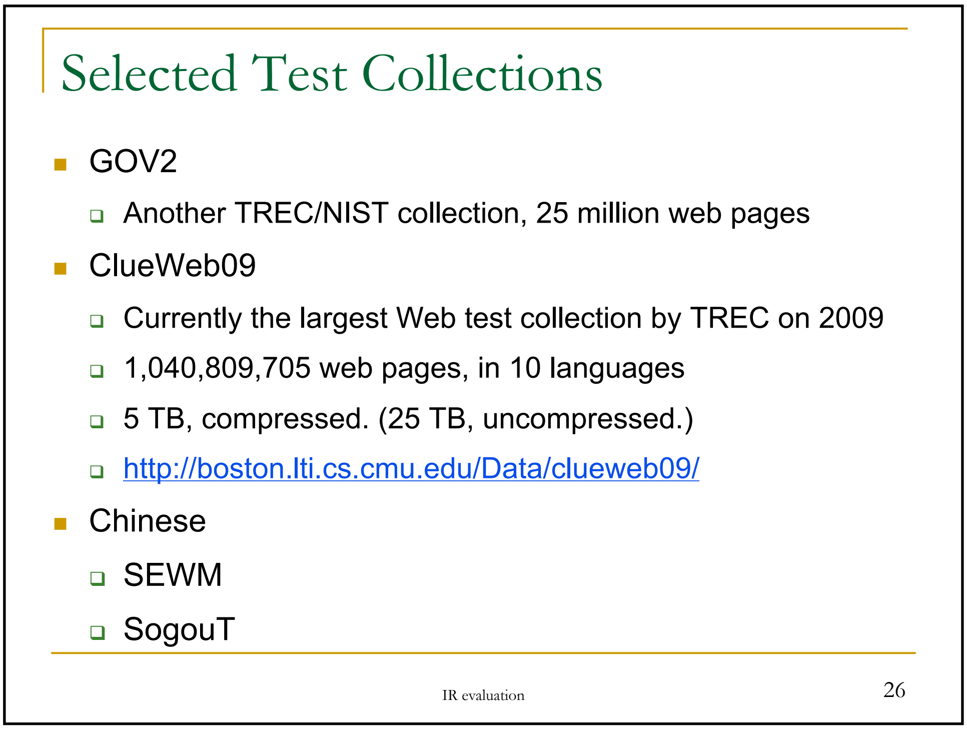
TREC : Pros and Cons

- Pros
  - Large-scale collections applied to common task
  - Allows for somewhat controlled comparisons
- Cons
  - Time-consuming in preparation and testing
  - Very long queries, also unrealistic
  - Comparisons still difficult to make, because systems are quite different on many dimensions
  - Also, topics used in every conference year present little overlap, which make the comparison difficult
  - Focus on batch ranking rather than interaction
    - There was an interactive track



Other Standard relevance benchmarks

- TRECVID
  - For Visual IR
- TAC
  - QA, Summarization, ……
- NTCIR
  - East Asian language and cross-language IR
- Cross Language Evaluation Forum (CLEF)
  - European languages and cross-language IR
- Many others
  - Yahoo! Challenges, Yandex Challenges, RecSys Challenges……



Selected Test Collections

- GOV2
  - Another TREC/NIST collection, 25 million web pages
- ClueWeb09
  - Currently the largest Web test collection by TREC on 2009
  - 1,040,809,705 web pages, in 10 languages
  - 5 TB, compressed. (25 TB, uncompressed.)
  - http://boston.lti.cs.cmu.edu/Data/clueweb09/
- Chinese
  - SEWM
  - SogouT

Other Test Collections (Cont.)-SogouT

- SogouT *SogouT 2019 is coming soon!*
  - Web pages: 138,700,000 Chinese Web pages, ~ 5 Terabyte
    - With link graph and SogouRank scores (an improve PageRank)
  - Query set
    - Most frequently requested **10,000** queries on Jun. 2008
    - 56% of the all the user queries (in terms of query frequency)
  - Annotation set
    - Automatically annotated **65,465** answers
    - 95% precision on the annotation, and 0.97 correlation with human annotation by sampled evaluation
  - The collection with the largest number of annotations
  - Have distributed 50+ copies to China, US, UK, CAN, JP.
  - Been used in NTCIR Intent-1&2 (2012,2013) Imine(2014)

Evaluation

Evaluation methodology

Standard benchmarks for IR

Annotation consistency

Common metrics for relevance

Inter-judge Agreement

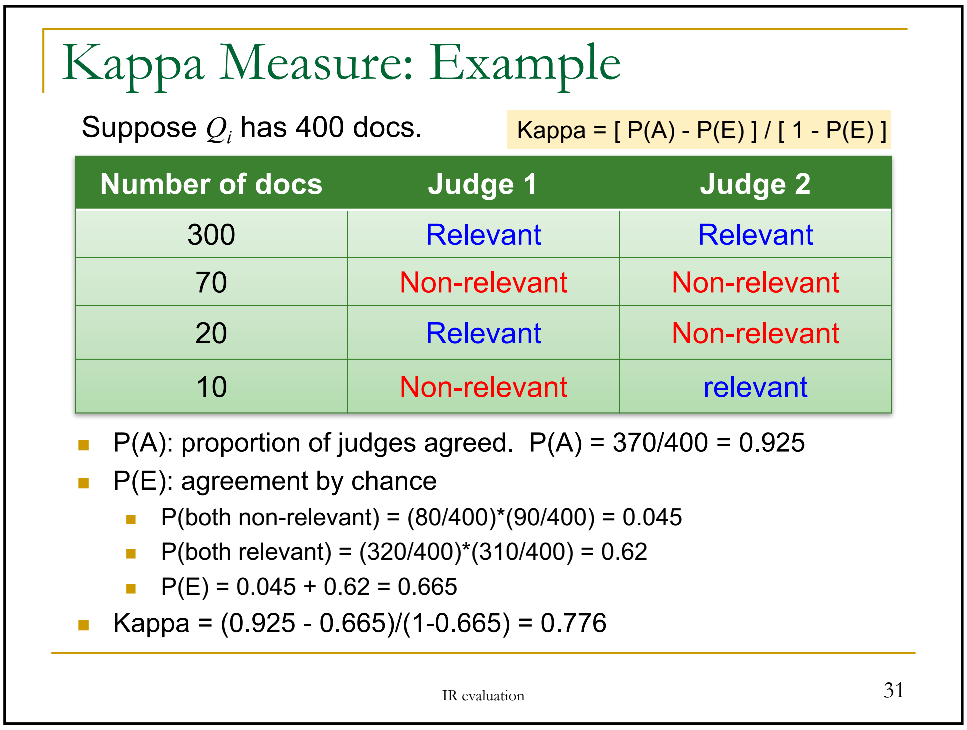
- Examples in TREC judgment

Topic	number of docs judged	disagreements	NR	R
51	211	6	4	2
62	400	157	149	8
67	400	68	37	31
95	400	110	108	2
127	400	106	12	94

NR(R): eventually classified as non-relevant (relevant)

Cohen’s Kappa coefficient measure for inter-judge (dis)agreement

- Kappa measure
  - Agreement measure among judges
  - Designed for categorical judgments
  - Corrects for chance agreement
- Kappa = [P(A) - P(E)] / [1 - P(E)] (Kohen, 1960)
  - P(A) : proportion of judges agreed
  - P(E) : what agreement would be by chance
  - For complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.



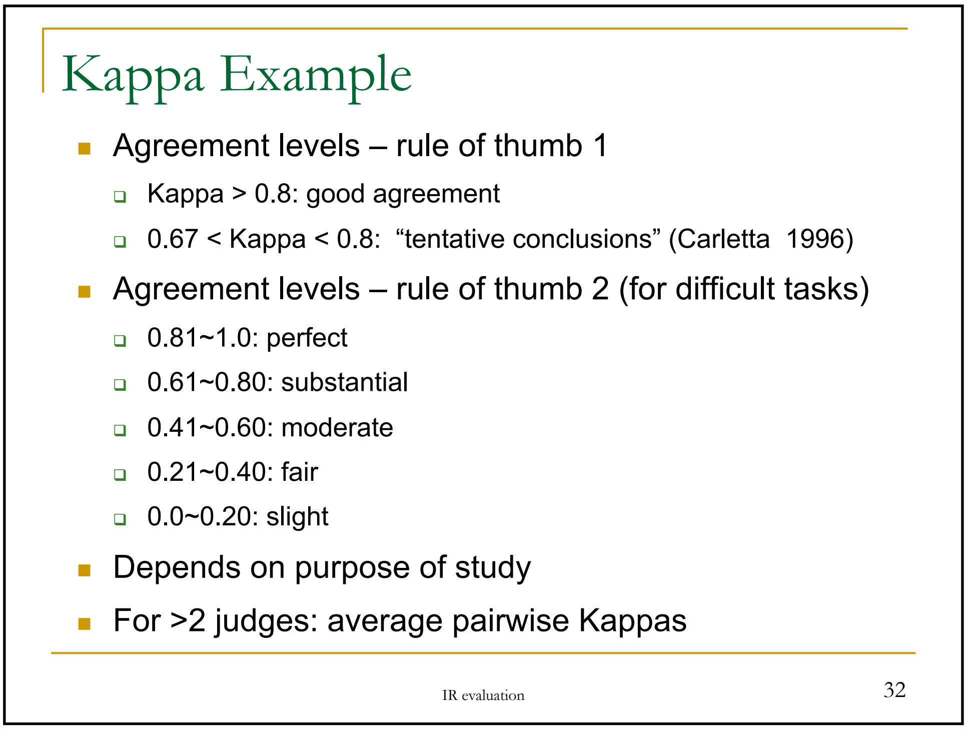
Kappa Measure: Example

Suppose Q_i has 400 docs.

Kappa = [P(A) - P(E)] / [1 - P(E)]

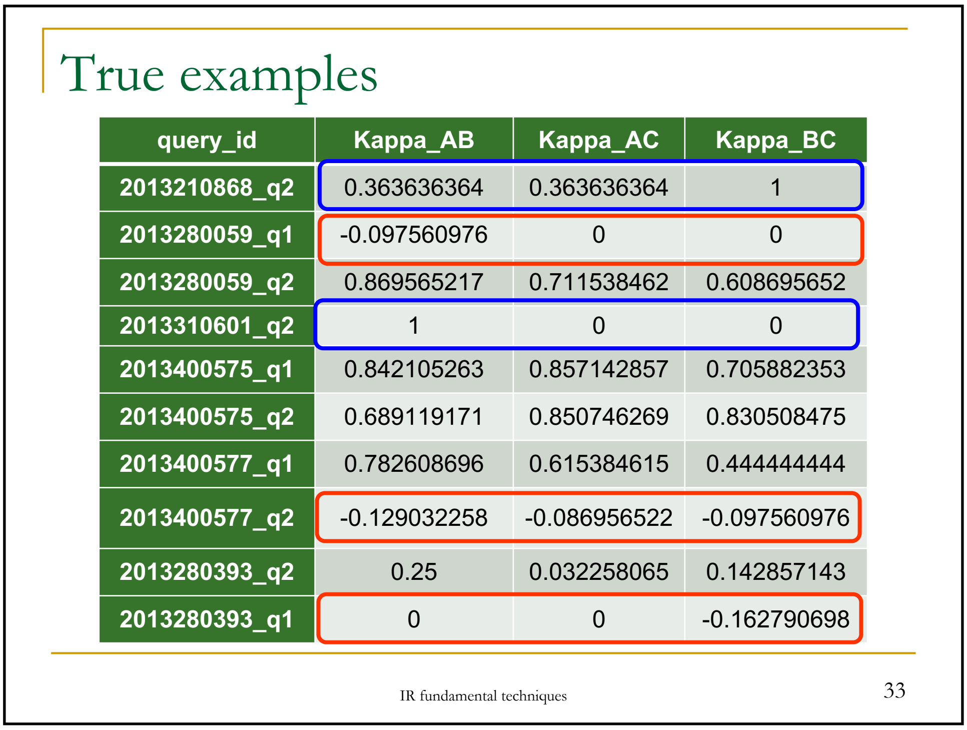
Number of docs	Judge 1	Judge 2
300	Relevant	Relevant
70	Non-relevant	Non-relevant
20	Relevant	Non-relevant
10	Non-relevant	relevant

- P(A): proportion of judges agreed. P(A) = 370/400 = 0.925
- P(E): agreement by chance
  - P(both non-relevant) = (80/400)*(90/400) = 0.045
  - P(both relevant) = (320/400)*(310/400) = 0.62
  - P(E) = 0.045 + 0.62 = 0.665
- Kappa = (0.925 - 0.665)/(1-0.665) = 0.776



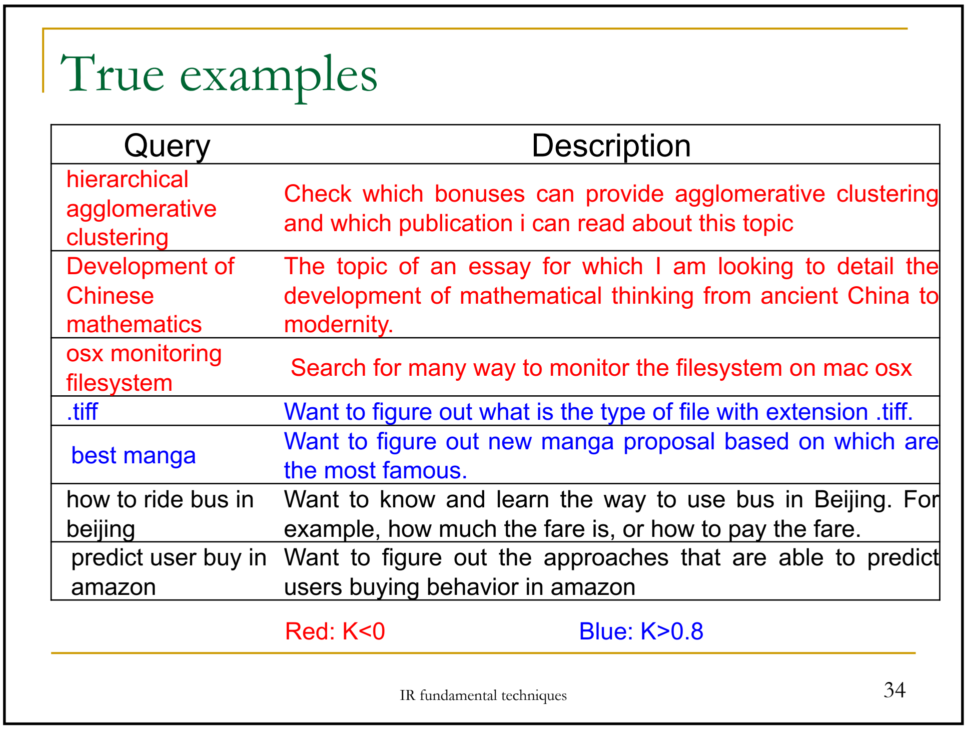
Kappa Example

- Agreement levels – rule of thumb 1
  - Kappa > 0.8: good agreement
  - 0.67 < Kappa < 0.8: “tentative conclusions” (Carletta 1996)
- Agreement levels – rule of thumb 2 (for difficult tasks)
  - 0.81~1.0: perfect
  - 0.61~0.80: substantial
  - 0.41~0.60: moderate
  - 0.21~0.40: fair
  - 0.0~0.20: slight
- Depends on purpose of study
- For >2 judges: average pairwise Kappas



True examples

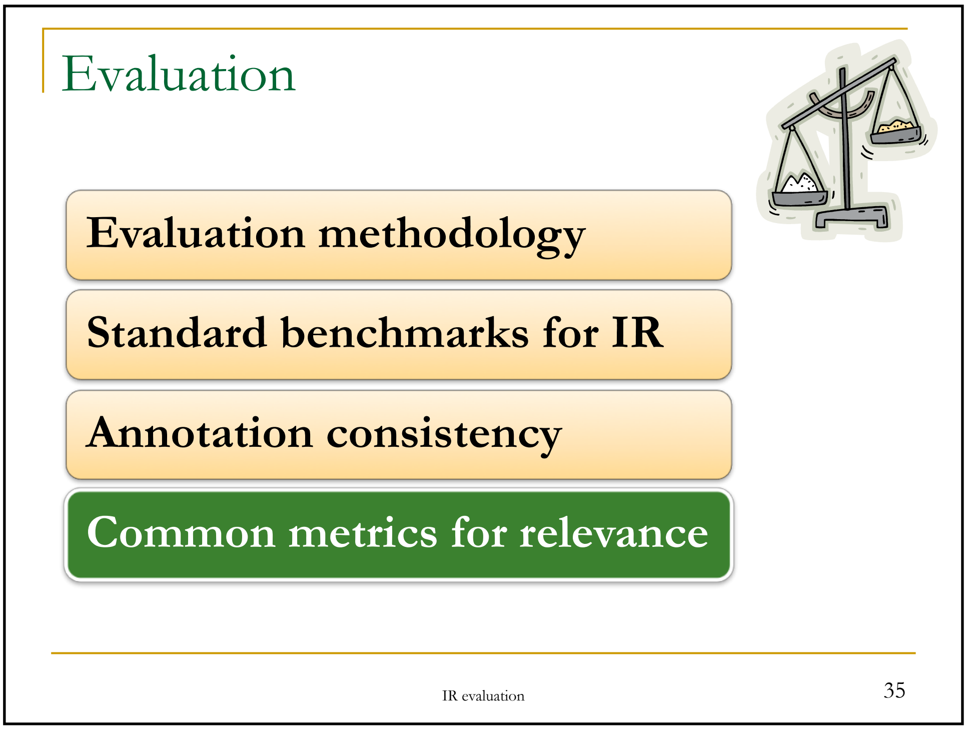
query_id	Kappa_AB	Kappa_AC	Kappa_BC
2013210868_q2	0.363636364	0.363636364	1
2013280059_q1	-0.097560976	0	0
2013280059_q2	0.869565217	0.711538462	0.608695652
2013310601_q2	1	0	0
2013400575_q1	0.842105263	0.857142857	0.705882353
2013400575_q2	0.689119171	0.850746269	0.830508475
2013400577_q1	0.782608696	0.615384615	0.444444444
2013400577_q2	-0.129032258	-0.086956522	-0.097560976
2013280393_q2	0.25	0.032258065	0.142857143
2013280393_q1	0	0	-0.162790698



True examples

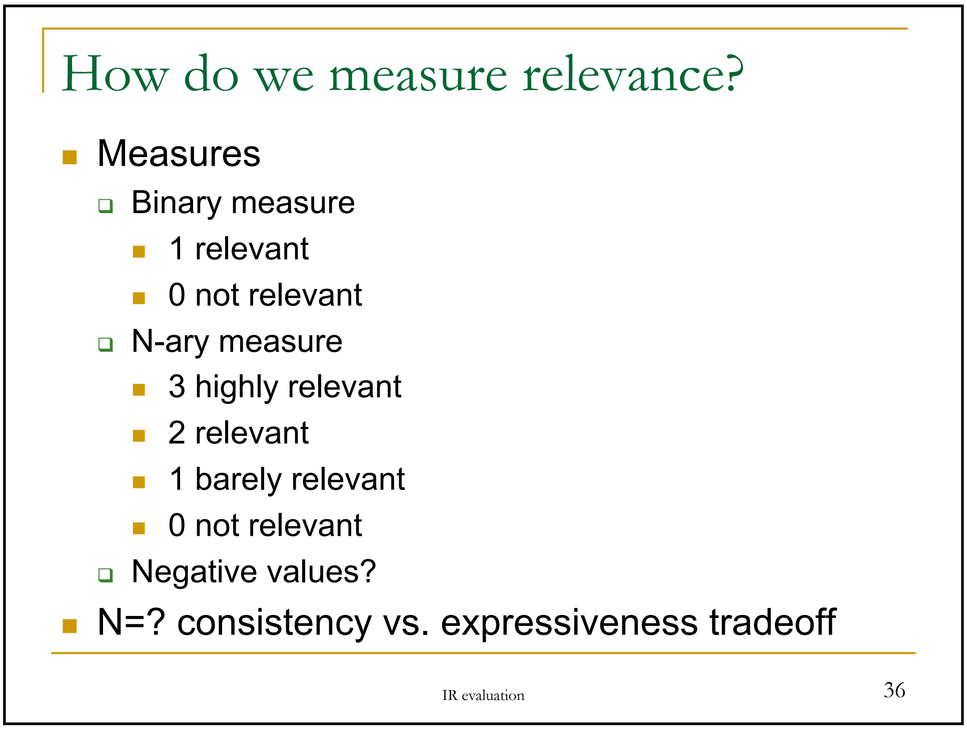
Query	Description
hierarchical agglomerative clustering	Check which bonuses can provide agglomerative clustering and which publication i can read about this topic
Development of Chinese mathematics	The topic of an essay for which I am looking to detail the development of mathematical thinking from ancient China to modernity.
osx monitoring filesystem	Search for many way to monitor the filesystem on mac osx
.tiff	Want to figure out what is the type of file with extension .tiff.
best manga	Want to figure out new manga proposal based on which are the most famous.
how to ride bus in beijing	Want to know and learn the way to use bus in Beijing. For example, how much the fare is, or how to pay the fare.
predict user buy in amazon	Want to figure out the approaches that are able to predict users buying behavior in amazon

Red: K<0 Blue: K>0.8



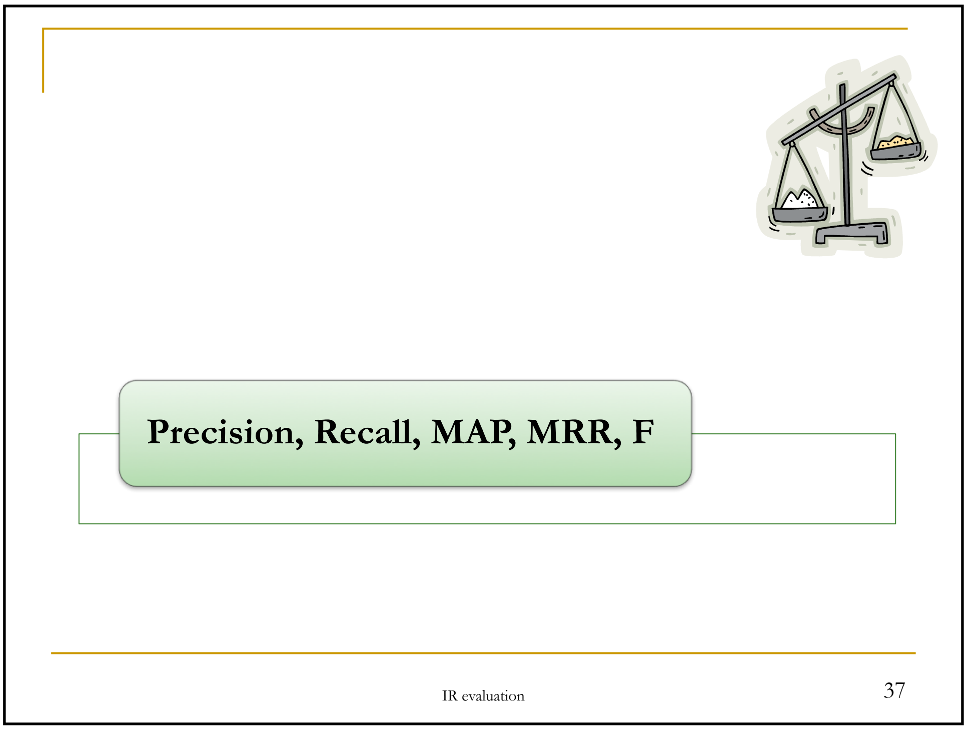
Evaluation

- Evaluation methodology
- Standard benchmarks for IR
- Annotation consistency
- **Common metrics for relevance**

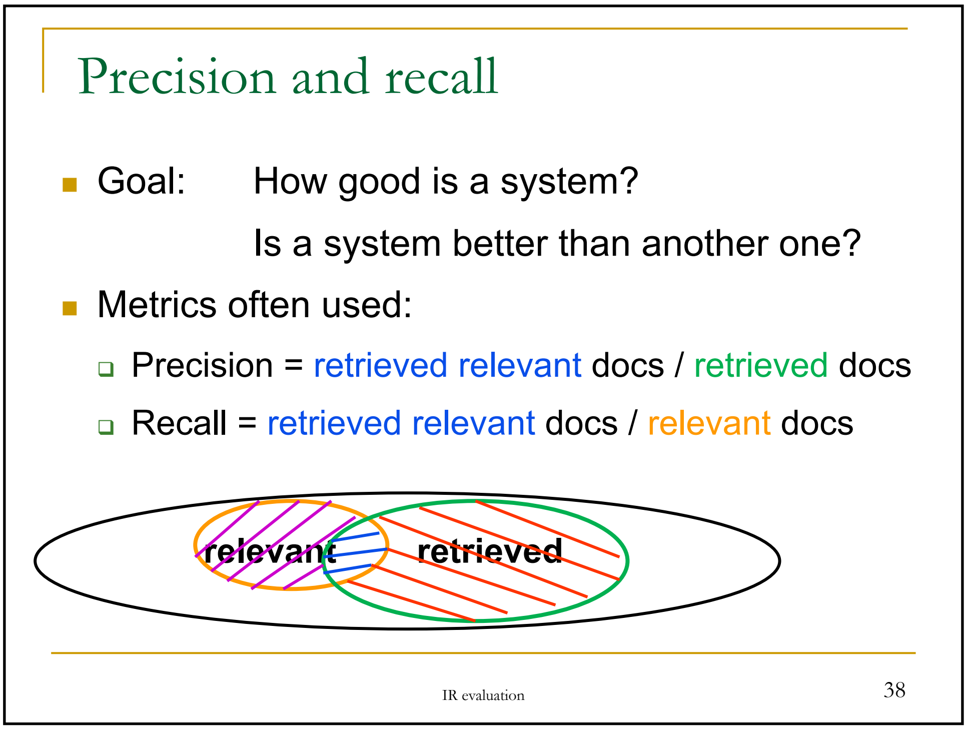


How do we measure relevance?

- Measures
  - Binary measure
    - 1 relevant
    - 0 not relevant
  - N-ary measure
    - 3 highly relevant
    - 2 relevant
    - 1 barely relevant
    - 0 not relevant
  - Negative values?
- N=? consistency vs. expressiveness tradeoff

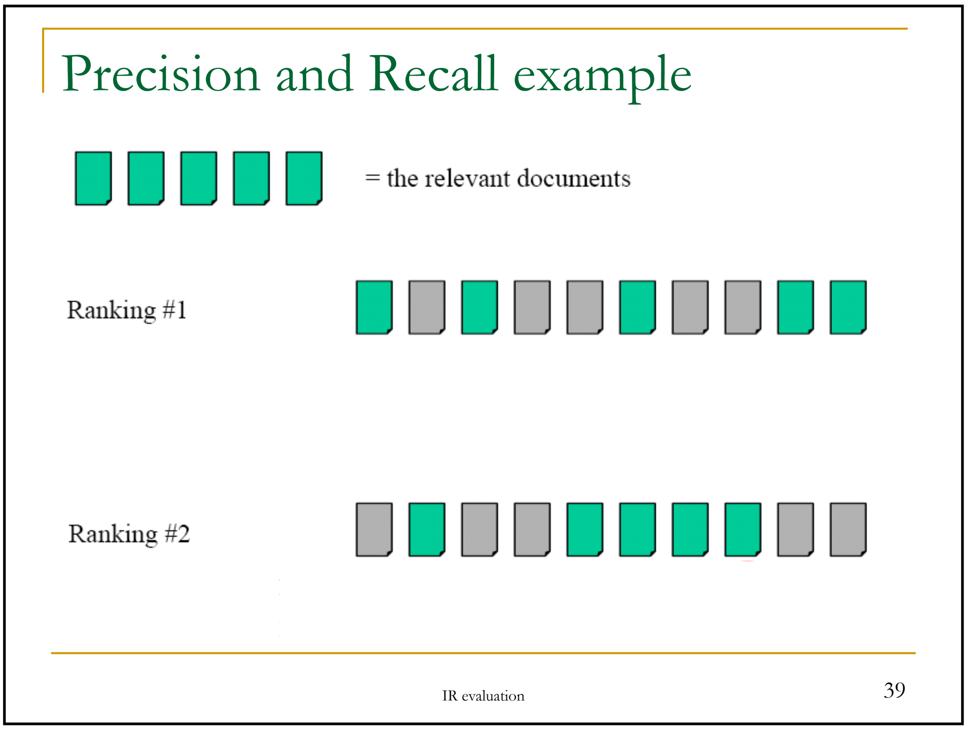


Precision, Recall, MAP, MRR, F



Precision and recall

- Goal: How good is a system?
  Is a system better than another one?
- Metrics often used:
  - Precision = retrieved relevant docs / retrieved docs
  - Recall = retrieved relevant docs / relevant docs

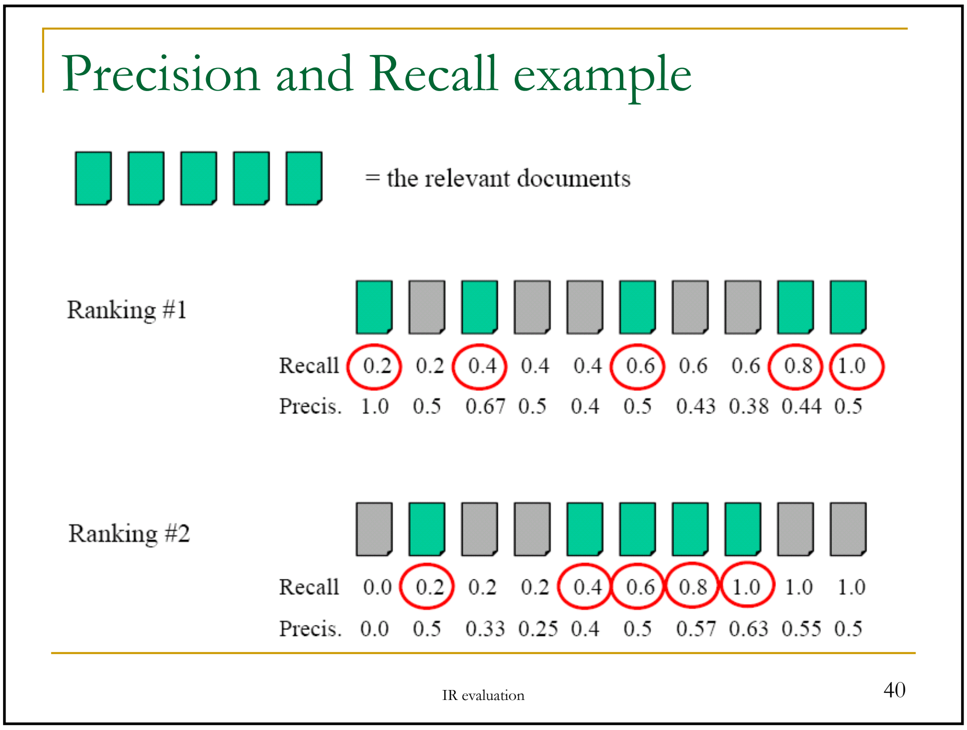


Precision and Recall example

🟩 🟩 🟩 🟩 🟩 = the relevant documents

Ranking #1 🟩 ⬜ 🟩 ⬜ ⬜ 🟩 ⬜ ⬜ 🟩 🟩

Ranking #2 ⬜ 🟩 ⬜ ⬜ 🟩 🟩 🟩 🟩 ⬜ ⬜



Precision and Recall example

🟩 🟩 🟩 🟩 🟩 = the relevant documents

Ranking #1

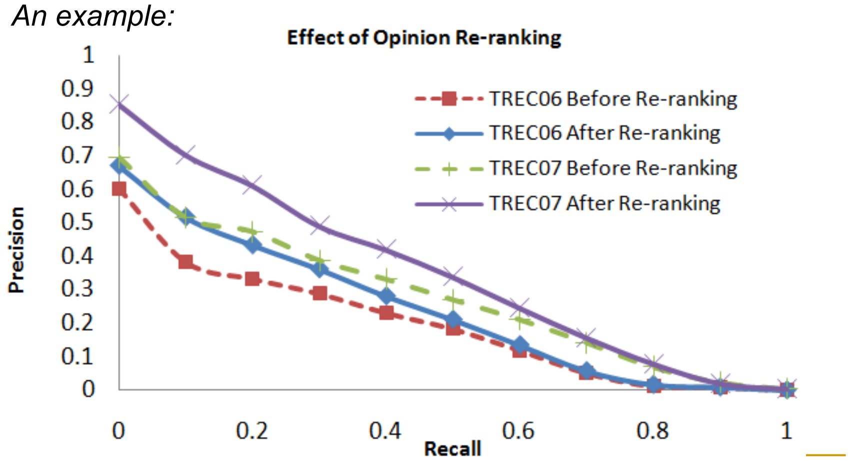
	🟩	⬜	🟩	⬜	⬜	🟩	⬜	⬜	🟩	🟩
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

Ranking #2

	⬜	🟩	⬜	⬜	🟩	🟩	🟩	🟩	⬜	⬜
Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precis.	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

Precision-recall curve

- By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*
- *An example:*



Summary measures

- People often want a single-number effectiveness measure (a summary measure)
- 1. Average precision is widely used in IR
  - Calculate by averaging precision when recall increases

$$AP = \frac{1}{N}\sum_{i=1}^{N} Pr\,ecision(i)$$

N is # of relevant results.
is NOT # of returned relevant results.

🟩⬜🟩⬜⬜🟩⬜⬜🟩🟩

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

AvgPrec= 62.2%

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

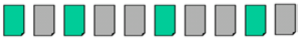
AvgPrec= 52.0%

Summary measures

- People often want a single-number effectiveness measure (a summary measure)
- 1. Average precision is widely used in IR
  - Calculate by averaging precision when recall increases

$$AP = \frac{1}{N}\sum_{i=1}^{N} Precision(i)$$

N is # of relevant results.
is NOT # of returned relevant results.



Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	0.8
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.40

AvgPrec= 2.61/5=52.2%

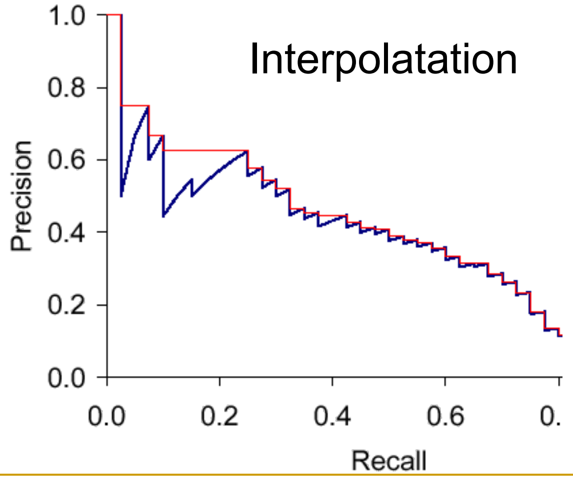
Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precis.	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

AvgPrec= 52.0%

Summary measures (cont.)

- 2. 11-point interpolated average precision
  - The standard measure in the early TREC
  - Take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents
  - Using interpolation (the value for 0 is always interpolated!)

11-point interpolated average precision (an example)



Summary measures (cont.)

- 2. 11-point interpolated average precision
  - The standard measure in the early TREC
  - Take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents
  - Using interpolation (the value for 0 is always interpolated!)
    - $p_{\text{interp}}(r) = \max p(r')$, where $r' >= r$ r, r': recall level

Original:

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

AP = 0.532 ?

recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precis.			1.0		0.67		0.5		0.44		0.5

Summary measures (cont.)

- 2. 11-point interpolated average precision
  - The standard measure in the early TREC
  - Take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents
  - Using interpolation (the value for 0 is always interpolated!)
    - $p_{\text{interp}}(r) = \max p(r')$, where $r' >= r$ r, r': recall level
  - And average them
  - Evaluates performance at all recall levels

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

AP=0.667

Final:

recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precis.	**1.0**	**1.0**	1.0	**0.67**	0.67	**0.5**	0.5	**0.5**	**0.5**	**0.5**	0.5

~~0.44~~

Summary measures (cont.)

- 3. Reciprocal Rank (RR)

$$RR = \frac{1}{Rank(first\_relevant\_result)}$$

- 4. Precision at top n document ($p@n$)
  - e.g. n =1,5,10

$$\Pr ecision@n = \Pr ecision(top\ n\ results)$$

Summary measures (cont.)

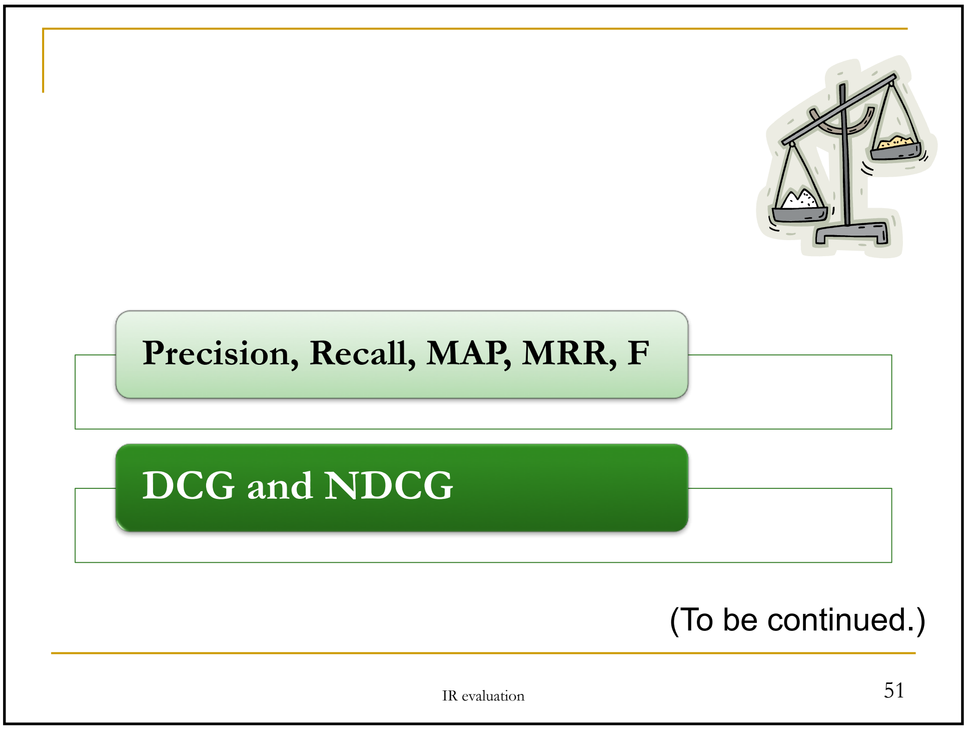
- 5. F-measure
  - Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

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

  - People usually use balanced F_1 measure
    - i.e., with β = 1 or α = ½
  - Harmonic mean is a conservative average
    - See CJ van Rijsbergen, *Information Retrieval*

Overall performance on multiple queries

- Using *Mean* value of evaluation scores on multiple topics/queries
  - MAP: Mean Average Precision
  - MAP(11-point): Mean 11-point AP
  - MRR: Mean Reciprocal Rank
  - ……



Precision, Recall, MAP, MRR, F

DCG and NDCG

(To be continued.)