Lecture 8 IIR Evaluation measures

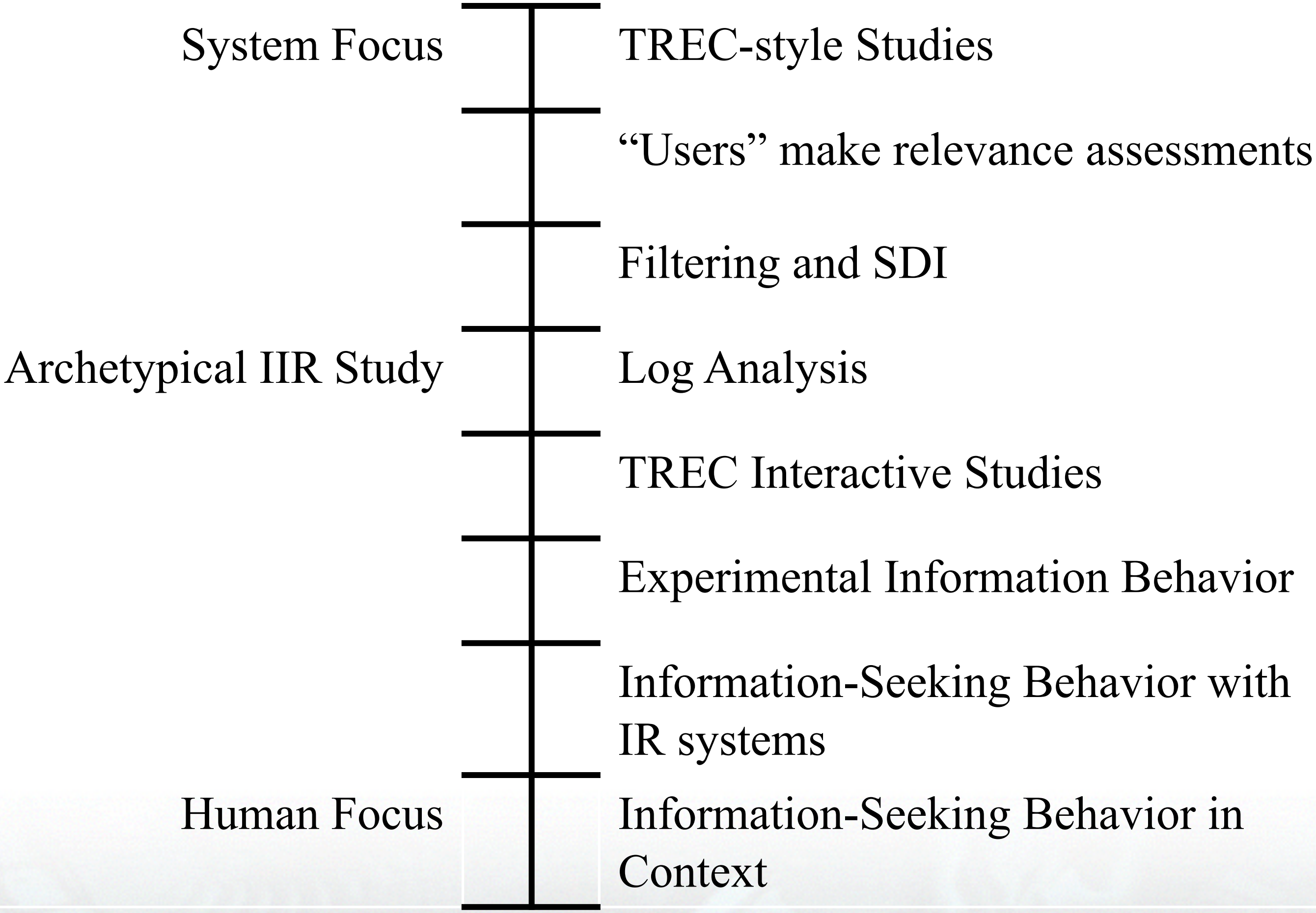
Chang Liu

刘畅

本节课内容参考资料

- Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1),
- Chapter 10: Measures, 99-125.

交互式信息检索
Interactive Information Retrieval



System Focus

Archetypical IIR Study

Human Focus

TREC-style Studies

“Users” make relevance assessments

Filtering and SDI

Log Analysis

TREC Interactive Studies

Experimental Information Behavior

Information-Seeking Behavior with IR systems

Information-Seeking Behavior in Context

(modified from Kelly, 2009, p.11)

Measures （度量）

1 Contextual（情境）

2 Interaction（交互）

3 Performance（绩效）

4 Usability（可用性）

Measures

- Differences in the understanding of usability in HCI and IIR
  - HCI： performance ~ usability
  - IIR： performance is distinct from usability
    - usability refers to self-report measures;
    - performance has always been an important evaluation measure in IR.

Definitions of measures

- Nominal vs. operational
  - Nominal definitions （名词性定义）
    - state the meaning of concepts;
  - Operational definitions （操作性定义）
    - specify precisely how a concept (and its dimensions) will be measured.

（e.g.） Independent Variables

- Cognitive style:
  - EFT(Embedded Figure Test)
    - FDs
    - FIs
- Online search experience
  - Questionnaire
    - Novice searchers
    - Experienced searchers

（e.g.）Dependent Variables

- Search performance
  - the average length of time spent for retrieving information
  - the average number of nodes visited for retrieving information
- Navigational style
  - the average number of times a navigation or search tool was chosen
  - the average number of layers consecutively traversed

Selection and Interpretation of Measures

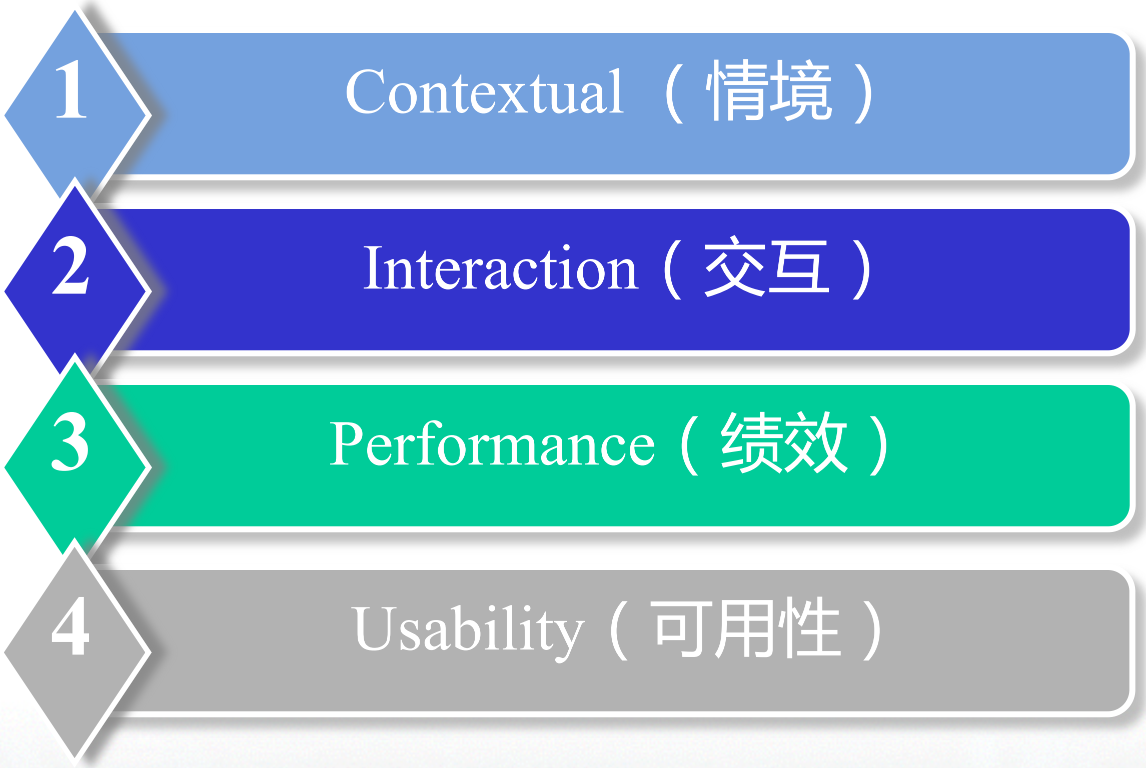
- Validity<有效性>
  - – the extent that the measurement procedures accurately reflect the concept we are studying.

- Reliability <可靠性>
  - – is demonstrated when measures are repeated under the same conditions and yield highly similar measurements each time

Selection and Interpretation of Measures

- The selection and interpretation of measures should be grounded by
  - the purposes of the system
  - the task the user is trying to accomplish
    - High-precision task or exploratory task

Measures （度量）



1 Contextual（情境）

2 Interaction（交互）

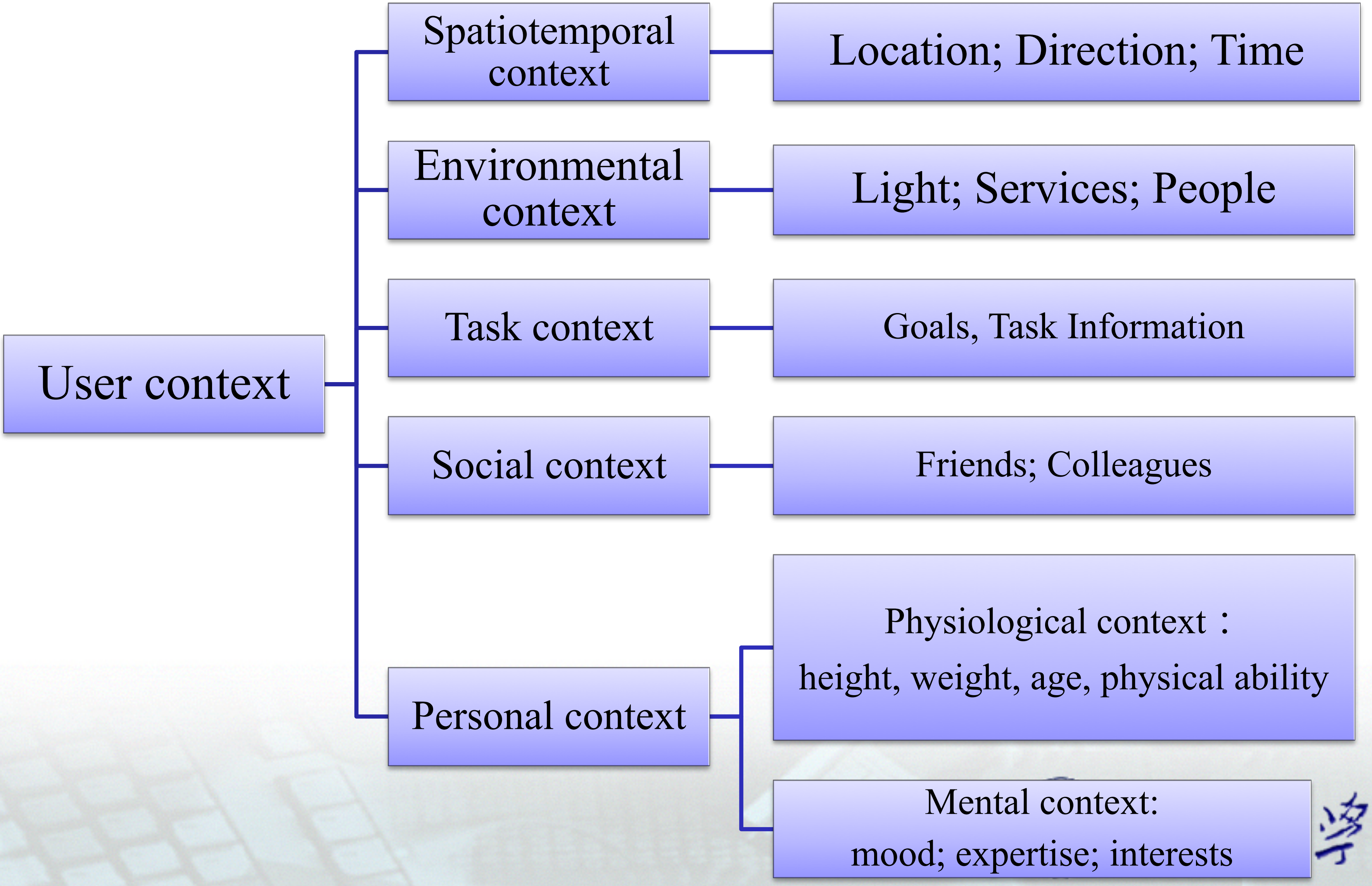
3 Performance（绩效）

4 Usability（可用性）

1. Context (情境的度量）

- Describe the context in which information search and interaction occurs

情境(Context)



- User context
  - Spatiotemporal context
    - Location; Direction; Time
  - Environmental context
    - Light; Services; People
  - Task context
    - Goals, Task Information
  - Social context
    - Friends; Colleagues
  - Personal context
    - Physiological context：height, weight, age, physical ability
    - Mental context: mood; expertise; interests

1. Context (情境的度量）

- Describe the context in which information search and interaction occurs
  - 1.1 Individual Differences
  - 1.2 Information Needs

1. Context

1.1 Individual Differences

- 可直接显示的用户特征(explicit features)
  - Gender, age, major, occupation, computer level, search skills, etc. 性别、年龄、专业、职业、计算机水平、搜索能力水平等
  - If the research goal is to examine these variables, then you should balance the sample according to the independent variables; 如果研究对象是这些变量：有目的的选取样本；
  - If the research goal is not to examine these variables, you can random select subjects, but you need to report their demographic features, and consider that as one of your limitations. 如果主要研究不是这些变量：也要汇报用户特征。
- 不可直接显示的用户特征 (implicit features)
  - Intellect, creativity, personality, memory, cognitive style ,cognitive ability, etc. 智力、创造力、个性、记忆力、认知风格等
  - 标准化仪器或工具(standardized instruments)

1. Context

1.2 Information Needs

- Task-related measures
  - task-type
  - task familiarity, task difficulty，topic familiarity and domain expertise
  - Difficulty: some of the measures are difficult to measure, e.g. domain expertise
  - 难点：很难精确地测量（如domain expertise）

2. Interaction

- Direct measures
  - number of queries
  - query length
  - number of search results viewed
  - number of documents viewed
  - number of documents saved
  - frequency counts of the activities
- Combined measures
  - time divided by the number of documents saved
  - the number of documents saved divided by the number of documents

Classification of interaction variables
交互变量列表

根据何时变量可以获取和计算 (According to when the variables are available and could be calculated)：

- 整体变量 Whole-session level variables
- 过程变量 Within-session level variables

根据交互发生的页面和地点(According to the location and the page type when interactions occur)：

- 查询式相关 Query-related
- 内容页面相关 Content－page related
- 搜索结果页面相关 SERP related
- 两两查询式间隔 Query interval related

交互变量列表

Whole-session level:

- Task completion time
- Numbers of all documents
- Numbers of unique documents
- Number of SERPs
- Number of unique SERPs
- Number of queries
- Total time spent on documents
- Total time spent on SERPs
- Ratio of document time to all
- Ratio of SERP time to all

Within-session level:

- Mean dwell time of all documents
- Mean dwell time of unique documents
- Mean dwell time of all SERPs
- Mean dwell time of unique SERPs
- Number of documents per query
- Number of unique documents per query
- Number of SERPs per query
- Number of unique SERPs per query
- Average query interval
- Average Time To First Click

2. Interaction

- 难点：如何解释这些交互变量？
- Difficulty: How to interpret the interaction variables?

Question：If a subject enters a large number of queries, is this good or bad?

Interaction variables

- Answer：For IR system evaluation, we should consider the goal of the design of information systems.
  - If the purpose of the system is to help a subject learn more about a topic, then more queries might be a positive indicator.
  - If the purpose of the system is to help a subject find a single answer, then more queries might be a negative indicator.

3. Performance

- Traditional IR evaluation measures (传统的信息检索系统评估指标)
  - Precision
  - Recall

- 隐藏的指标：Relevance （相关度）
  - Relevance is often considered as binary, static, unidimensional and generalizable.

Relevance as criterion for measures

Precision

- Probability that what is retrieved is relevant
  - conversely: how much junk is retrieved?

Recall

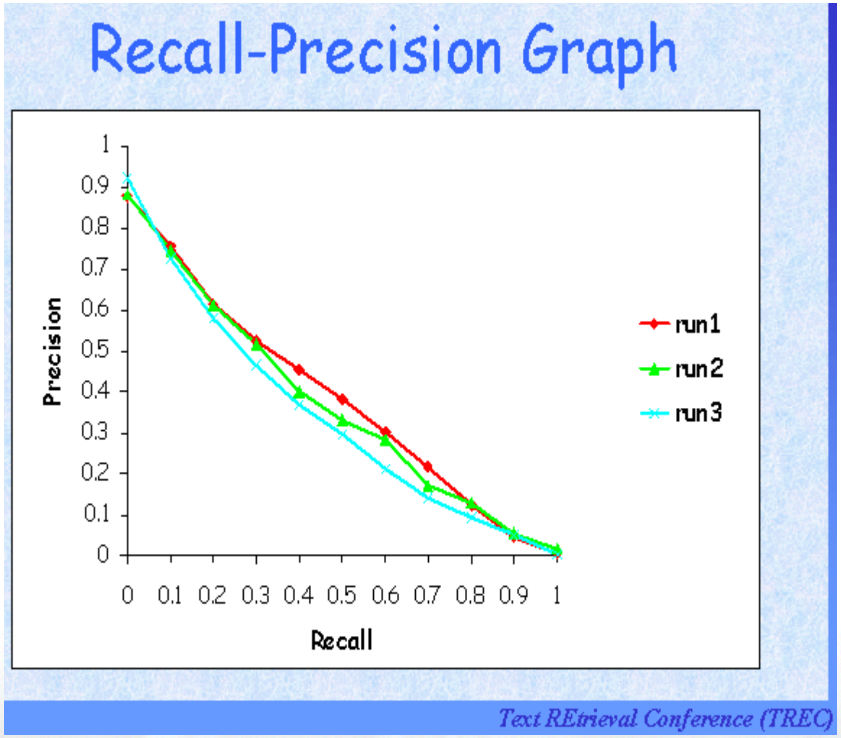
- Probability that what is relevant in a file is retrieved
  - conversely: how much relevant stuff is missed?

- Probability of agreement between what the system retrieved/not retrieved as relevant (*systems relevance*) & what the user assessed as relevant (*user relevance*) where user relevance is the gold standard for comparison

Tradeoff in recall vs. precision

- Generally, there is a tradeoff:
  - recall can be increased by retrieving more but precision decreases
  - precision can be increased by being more specific but recall decreases
- Some users want high precision; others high recall

- Cleverdon’s law


Recall-Precision Graph

Text REtrieval Conference (TREC)

Assumptions in Cranfield methodology

- IR and thus relevance is static (traditional IR model)
- Further: Relevance is:
  - topical
  - binary
  - independent
  - stable
  - consistent
  - if pooling: complete

- Inspired relevance experimentation on every one of these assumptions
- Main finding: none of them holds

but these simplified assumptions enabled rich IR tests and many improvements

IR & relevance: static vs. dynamic

Q: Do relevance inferences & criteria change over time for the same user & task?

A: They do

- For a given task, user's inferences are dependent on the stage of the task

IR & relevance inferences are highly dynamic processes

3. Performance

3.1 Traditional IR Performance Measures

3.2 Interactive Recall and Precision

3.3 Measures that Accommodate Multi-Level Relevance and Rank

3.4 Time-Based Measures

3.5 Informativeness

3.6 Cost and Utility Measures

3.1 Traditional IR Performance Measures

Measure	Description
Recall	The number of retrieved relevant documents divided by the number of relevant documents in corpus.
Precision	The number of relevant retrieved documents divided by the number of retrieved documents.
F-measure (F-score)	The F-measure is a way of combining precision and recall and is equal to their weighted harmonic mean [F = (β^2+1)(precision × recall)/(β^2precision + recall)]. The F-measure also accommodates weighting of precision or recall, to indicate importance.

3.1 Traditional IR Performance Measures

Measure	Description
Average precision (AP)	Individual precision scores are computed for each retrieved relevant document. These values are then summed and divided by the total number of relevant documents in the collection.
Mean average precision (MAP)	This is a run level measure and consists of taking the average of the average precision values for each topic.

3.1 Traditional IR Performance Measures

Measure	Description
Precision at n	The number of relevant documents in the top n results divided by n.
Mean reciprocal rank (MRR)	This measure was developed for high-precision tasks where only one or a small number of relevant documents are needed. For a single task with one relevant document, reciprocal rank is the inverse of its ranked position. MRR is the average of two or more reciprocal rank scores (used when there is more than one task).

3.2 Interactive Recall and Precision

- In IIR evaluations,
  - Subjects do not agree with the assessor’s relevance judgments
  - Subjects usually are unable to search through 1,000 documents.

3.2 Interactive Recall and Precision

Measure	Description
Interactive recall	Number of TREC relevant saved by user/number of TREC relevant documents in the corpus.
Interactive TREC precision	Number of TREC relevant documents viewed by the user/total number viewed.
Interactive user precision	Number of TREC relevant documents saved by the user/total number saved by the user.
Relative relevance (RR)	Cosine similarity measure between two lists of relevance assessments for the same documents

Veerasamy and Belkin (1996) and Veerasamy and Heikes (1997)

3.3 Measures that Accommodate Multi-Level Relevance and Rank

- Two other problems with traditional performance measures
  - binary relevance assessments
  - relevant documents that are retrieved further down on the results list are less useful

3.3 Measures that Accommodate Multi-Level Relevance and Rank

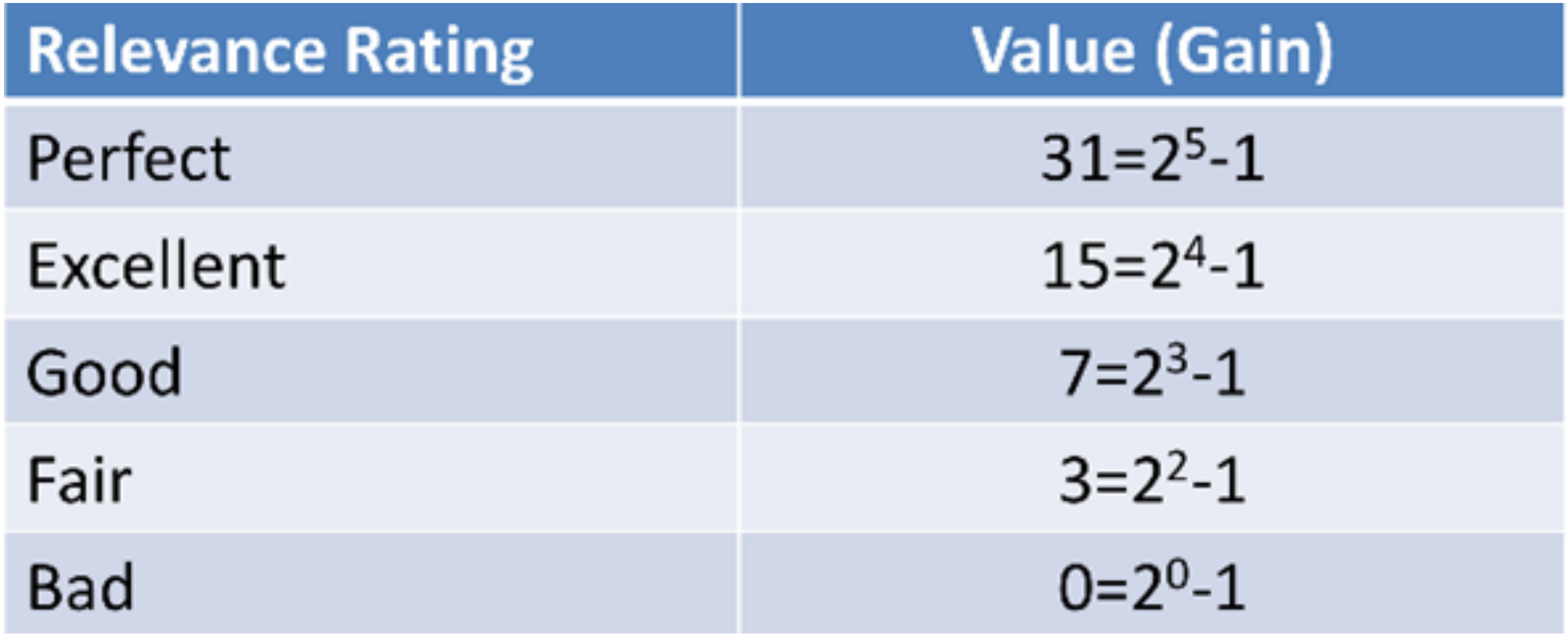
Measure	Description
Cumulated gain (CG)	Cumulated gain can be computed at different cut-off values for search result of lists of varying sizes. At the cut-off point, CG is the sum of the relevance values of all documents up to and including the document at the cut-off point.
Discounted cumulated gain (DCG)	DCG discounts the value of relevant documents according to their ranked position. Computed by dividing the relevance score of a document by the logarithm of its rank. The discounted relevance scores are then summed to a particular cut-off point.
Normalized discounted cumulated gain (nDCG)	The DCG measure is normalized according to the best DCG available for a given results list.
Ranked half-life (RHL)	The point in the results list at which half of the total relevance value for the entire list of documents has been achieved.

3.3 Measures that Accommodate Multi-Level Relevance and Rank

- Two assumptions of DCG:
  - Highly relevant documents are more useful than marginally relevant document
  - The lower a document's rank in a results list, the less likely the subject is to view it.

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

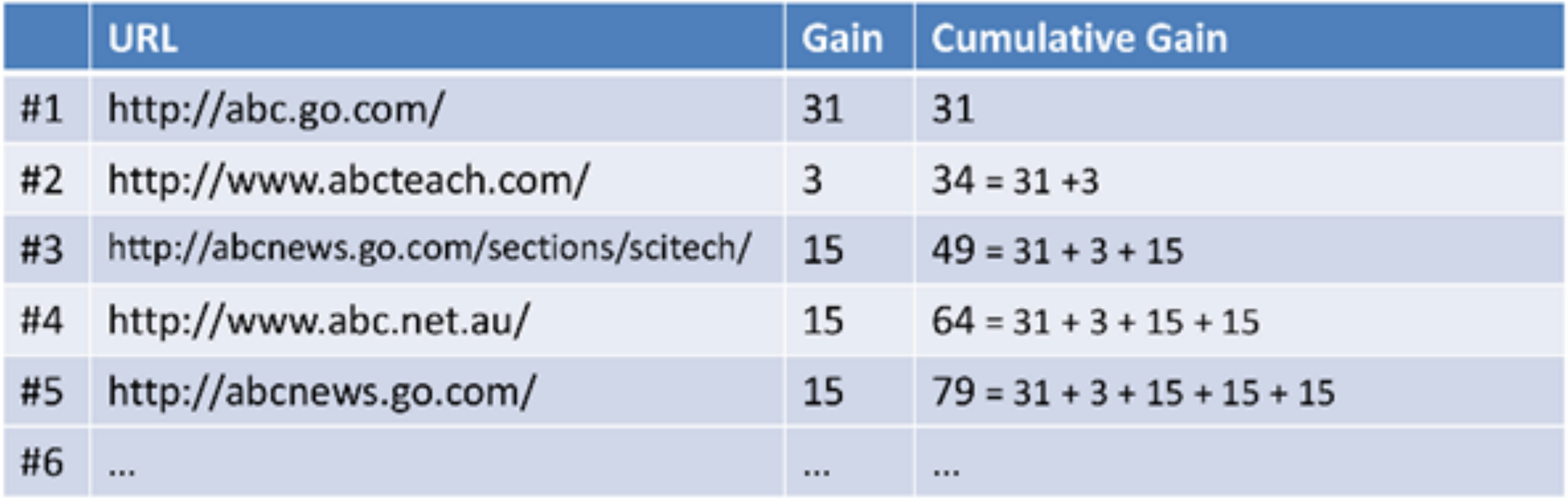
Discounted cumulated gain (DCG)



Relevance Rating	Value (Gain)
Perfect	$31=2^5-1$
Excellent	$15=2^4-1$
Good	$7=2^3-1$
Fair	$3=2^2-1$
Bad	$0=2^0-1$

Discounted cumulated gain (DCG)

Query={abc}



	URL	Gain	Cumulative Gain
#1	http://abc.go.com/	31	31
#2	http://www.abcteach.com/	3	34 = 31 +3
#3	http://abcnews.go.com/sections/scitech/	15	49 = 31 + 3 + 15
#4	http://www.abc.net.au/	15	64 = 31 + 3 + 15 + 15
#5	http://abcnews.go.com/	15	79 = 31 + 3 + 15 + 15 + 15
#6	…	…	…

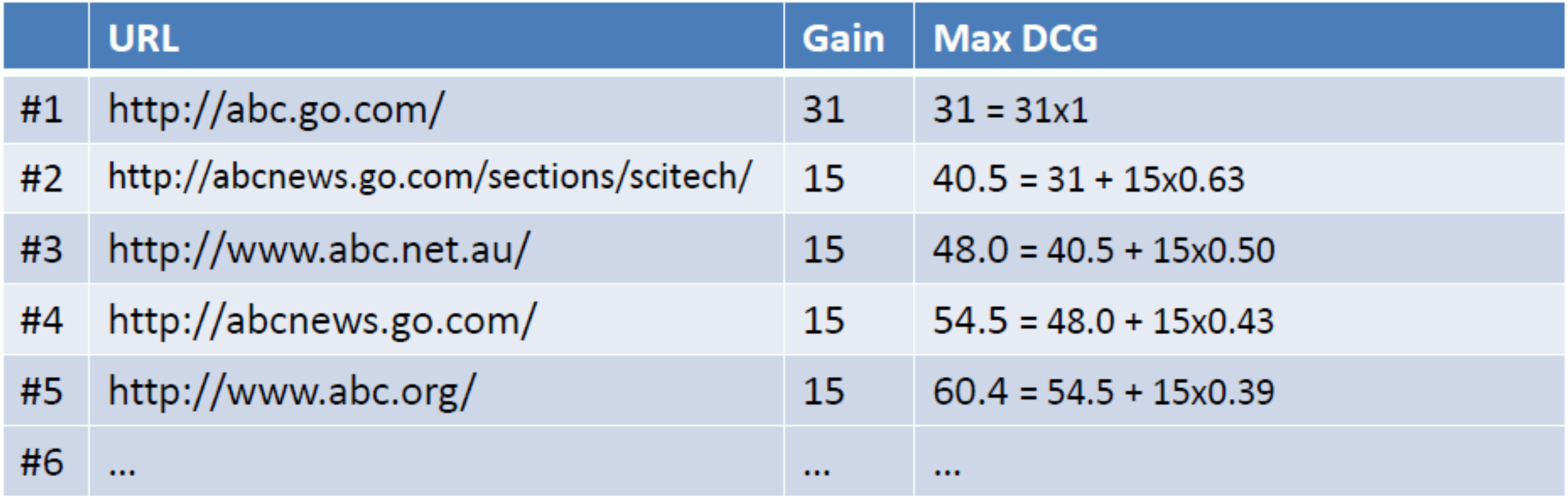
Discounted cumulated gain (DCG)

Discounting factor: log(2)/log(1+rank)

	URL	Gain	Discounted Cumulative Gain
#1	http://abc.go.com/	31	31 = 31x1
#2	http://www.abcteach.com/	3	32.9 = 31 + 3x0.63
#3	http://abcnews.go.com/sections/scitech/	15	40.4 = 32.9 + 15x0.50
#4	http://www.abc.net.au/	15	46.9 = 40.4 + 15x0.43
#5	http://abcnews.go.com/	15	52.7 = 46.9 + 15x0.39
#6	…	…	…

Discounted cumulated gain (DCG)

IDCG（ideal DCG），理想的DCG



	URL	Gain	Max DCG
#1	http://abc.go.com/	31	31 = 31x1
#2	http://abcnews.go.com/sections/scitech/	15	40.5 = 31 + 15x0.63
#3	http://www.abc.net.au/	15	48.0 = 40.5 + 15x0.50
#4	http://abcnews.go.com/	15	54.5 = 48.0 + 15x0.43
#5	http://www.abc.org/	15	60.4 = 54.5 + 15x0.39
#6	…	…	…

Normalized Discounted Cumulated Gain (nDCG)

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

	URL	Gain	DCG	Max DCG	NDCG
#1	http://abc.go.com/	31	31	31	1 = 31/31
#2	http://www.abcteach.com/	3	32.9	40.5	0.81=32.9/40.5
#3	http://abcnews.go.com/sections/scitech/	15	40.4	48.0	0.84=40.4/48.0
#4	http://www.abc.net.au/	15	46.9	54.5	0.86=46.9/54.5
#5	http://abcnews.go.com/	15	52.7	60.4	0.87=52.7/60.4
#6	…	…	…	…	…

3.4 Time-Based Measures

- Time-Based Measures have been used quite a lot in IIR evaluations
  - – a gross level
    - the length of time it takes a subject to complete a search task
  - – a more specific level
    - the length of time a subject spends viewing a search result or engaging in a specific action

3.4 Time-Based Measures

Measure	Description
Search speed	The proportion of answers that are found per minute. This measure consists of dividing the total number of answers found by the length of time it took to find the answers. All answers are included in this computation regardless of whether they are correct.
Qualified search speed	This measure accommodates multi-level relevance and consists of computing search speed for each relevance category, including non-relevant.

Kaki and Aula (2008)

3.5 Informativeness

- Absolute measures
- Relative evaluations of relevance
  - – Proposed by Tague (1998)
  - – The assumption behind this is that asking subjects to rank a set of search results from most informative to least informative results in more accurate data than asking them to associate absolute judgments with each result using a scale.
  - – None large-scale validation

3.6 Cost and Utility Measures

- In the early days, cost and utility measures figured prominently in the IR evaluation framework
  - It has been an important part of the evaluation of library and information services
- Nowadays, information is freely available online, so they are less relevant to the individual users

4. Evaluative Feedback from Subjects

4.1 Usability

4.2 Preference

4.3 Mental Effort and Cognitive Load

4.4 Flow and Engagement

4.5 Subjective Duration Assessment

4.6 Learning and Cognitive Transformation

4. Evaluative Feedback from Subjects

- 4.1 Usability

  “*to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction*” (ISO 1998)

  – Effectiveness: accuracy and completeness with which users achieve specified goals

  – Efficiency: resources expended in relation to the accuracy and completeness with which users achieve goals

  – Satisfaction: freedom from discomfort, and positive attitudes of the user to the product

4. Evaluative Feedback from Subjects

- 4.1 Usability
  - Effectiveness
  - Efficiency
  - Satisfaction
  - Ease of Use, Ease of Learning, and Usefulness

4.1 Usability

- Effectiveness
  - Precision and recall
  - Self-report data from subjects about their perceptions of performance
  - Completeness
  - Precision: ratio between correct information and total information retrieved
  - Recall: subjects' ability to recall information from the interface
  - Error rate

4.1 Usability

- Efficiency
  - – the time it takes a subject to complete a task
  - – the amount of time subjects spend doing different things or in different modes
- Satisfaction
  - – attempts to gauge subjects’ feelings about their interactions with the system
  - – how satisfied are you with your performance?

4.1 Usability

- Ease of Use: *the amount of effort which subjects expend executing and/or accomplishing particular tasks.*
- Ease of Learning: *how hard a system is to learn to use*
- Usefulness: *whether a tool is appropriate to the tasks and needs of the target users*

4.1 Usability

Available instruments for measuring usability

- Questionnaire for User Interface Satisfaction (QUIS) [53]
  - the subject’s overall reactions to the software, the screen, the terminology and system information, and learning and system capabilities.
- The USE questionnaire [188]
  - usefulness, ease of use, ease of learning, and satisfaction.
- Software Usability Measurement Inventory (SUMI) [256]

4. Evaluative Feedback from Subjects

4.2 Preference

In studies of two or more systems with a within-subjects design, it is common to collect preference information from subjects.

4. Evaluative Feedback from Subjects

4.3 Mental Effort and Cognitive Load

Mental demand, physical demand, temporal demand, performance, frustration and effort.

- NASA-Task Load Index (NASA-TLX)
- auxiliary tasks + primary task

4. Evaluative Feedback from Subjects

4.4 Flow and Engagement

Flow: “mental state of operation in which a person is fully immersed in what he she is doing, characterized by a feeling of energized focus, full involvement, and success in the process of the activity.” --Csikszentmihalyi

Engagement: “a quality of user experiences with technology that is characterized by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, and interest and affect”--O’Brien and Toms

4. Evaluative Feedback from Subjects

4.5 Subjective Duration Assessment

- Subjects estimate the length of time it took them to complete tasks, then compared to the actual length of time it took them.

underestimated times → high success rates

overestimated times → low success rates.

Czerwinski et al.

4. Evaluative Feedback from Subjects

4.6 Learning and Cognitive Transformation

- Evaluate how much a person knows is not easy
- Assess final products

第二次作业

- 假设你是一个搜索领域的专家，一份IT杂志邀请你写一篇有关搜索系统评估的文章。请你选择一个特定领域中的两个的搜索系统（注：不包含如Google、百度、Bing等一般搜索引擎），然后对这两个搜索系统进行对比评估。
- 文章要写的有深度、有说服力、有趣、有创造性。这个作业的目的是将课上学习到的交互式信息检索中的几个重要概念应用到实践中：如情境、搜索任务、评价指标、搜索体验等。
- 提交第二次作业,截止日期:2020/4/20.

第二次作业

- 首先，你需要选取一个使用搜索系统的情境，即具体的描述一种情境下人们会使用这两个搜索系统。
- 然后设想两个可能会应用这两个搜索系统完成的搜索任务：一个是简单的搜索任务，一个是复杂的搜索任务，将它们具体的描述出来。
- 你在这两个搜索系统上分别对两个搜索任务进行搜索，记录搜索过程和搜索体验，进而对两个系统进行对比分析。

第二次作业

- 对比分析中，请选取多个指标（至少五个）来描述搜索过程、系统搜索性能和用户搜索体验。需要强调的是不要仅仅关注搜索结果本身，更要关注用户的搜索体验，如搜索结果是否简单易懂？搜索过程中是否有迷失的情况？用户在搜索过程中可能会遇到什么困难？

- 导言
  - 介绍所选择的领域和两个搜索系统，并阐述这个领域的重要性，以及你选取这两个搜索系统的原因。在描述具体系统的时候，最好给出两个系统的网址和主页的网页快照。
- 搜索任务和评估指标
  - 对你选取的两个搜索任务进行具体描述，然后说明评估过程中所选取的指标，及指标选取的依据。
- 评估结果
  - 根据你的搜索任务和选取的评估指标对两个系统的搜索性能和搜索体验进行对比分析。在分析中，建议使用图表等方式总结进行对比，对较重要的发现要有强调，展示结果的方式可以自由发挥。
- 总结及对系统建议
  - 总结这两个系统的优缺点，并且提出，在你选取的领域中，考虑到用户的背景、搜索任务和不同的情境，你更推荐哪个系统？

第二次作业

- 评估指标的选取：
  - 注意recall 查全率的计算问题
  - DCG或nDCG的计算问题 可以选用@10
  - 不能全部是用户主观评价的指标

- 最后要有系统的综合比较
- 系统建议要与之前的评估指标对应

本周思考题-1

- 你还知道哪些信息检索评估的指标？可以跟我们分享出来，并给出相应的参考文献。

- If you know other evaluation measures for Information retrieval systems, please share them with us. It is recommended to include the reference.

本周思考题-2

- 你每天休闲的时候浏览的信息系统或app有哪些？
- 你父母或家里的长辈每天经常浏览的app或信息源有哪些？
- 你了解到的家里亲戚的中小学生休闲时候经常查看和浏览的app或信息源有哪些？
- What Apps do you usually browse during leisure time? What about your parents or elderly people in your family? How about the teens? Please share these apps in leisure time with us.