Indexing and Abstracting

Lecture 07 -- Evaluation of Indexing

Kuang-hua Chen
Department of Library and Information Science
National Taiwan University

Evaluation of Indexing

- Effectiveness
- Efficiency
- Completeness

- A good index is the result of many factors

Information is Critical

- Unretrieved information is the same as nonexistent information
- Factors make impacts on information retrieval system
  - Indexing
  - File structuring
  - Coding
  - Faulty searching procedures
  - Bad computer programming
  - User interface
  - …
- System factors and human factors affect indexing

System Factors

- Nature of index language
- Constraints of exhaustivity
- Constraints of specificity
- Level of coordination
- Overall structure

Human Factors

- Indexing consistency
- Subject expertise
- Indexing accuracy
- Indexing experience

Close-up to Human Factors

- Input and output to an information retrieval system
- Input part
  - Consistency among multiple indexers
  - Experience of indexers
  - Subject-matter knowledge of indexers
- Output part
  - Experience of searchers
  - Skill of searchers
  - Subject-matter knowledge of searchers

Evaluation

- A system with theoretical body could be evaluated in a quantitatively way
- Each step in indexing are similarly critical
- Too many variables cannot be dealt with in indexing evaluation together
- A controlled environment is built for indexing evaluation

Evaluation – The general problem

- What is a good index?
- Define goodness in terms of objectives
  - Does it fulfill its stated purposes?
  - Are its scope and coverage adequate?
  - Does it meet information need of users?
- Indicators
  - Accuracy
  - Consistency
  - Form
  - Internal structure

Approaches

- Evaluation of a single index
  - Needs of clientele
  - Subjects covered
  - Stated purposes
  - cost
- Comparison of multiple indexes
  - Relative quality
  - Relative cost

Indexing Comparisons have made

- Human indexing has been inter-compared for consistency
- Human indexing has been compared with machine indexing
- Relative utility of using different parts of a document for indexing
- Statistical methods and quasi-mathematical models have been proposed to ascertain quality of indexes

The Problem

- Subjective nature of what a *good* index is

Milestone of Indexing Evaluation

- Cranfield I
  - Focus on indexing and searching
  - Simple model
    - Collect a set of test documents
    - Devise a search procedure
    - Submit artificial queries
    - Judge the relevance of retrieved documents
  - If the results were poor, the fault was attributed to the indexing
- Cranfield II

Controversy Results of Cranfield II

- Simple term index language give better results
- Groups of terms drop in retrieval performance while single term index language used
- Simple coordination gives better precision than more complex devices

- Simple is the best?
- Still debate

Evaluation based on User's Need

- The user's external expression of need may not truly express the internal need
- Users know what is needed but do not realize that they are not expressing it the way that the system requires

- *Before evaluation can be carried out, some criteria of user needs and demands concerning an index must be established*

Types of User Needs

- Overt information related to the item
  - Author or title
- A subject need that is specific and well-defined
- A vague and ill-defined need

Relevance

- Indexing evaluation will never be effective until there is an understanding of the percept of relevance
- The search result against a query is to separate the all documents into two parts
  - One part is the set of relevant documents
  - The other part is the set of irrelevant documents

Relevance and Pertinence

- Relevance is the relationship between a document and a request
- Relevance is associated with the relationship between document and index
- Pertinence is the relationship between a document and a user
- Pertinence is concerned with the immediate usefulness to a particular user

Relevance and Pertinence (continued)

- Documents are relevant to query but not pertinent to the user
  - Documents are not timely
  - Documents are in foreign languages
  - Documents are beyond the understanding of the user
  - Documents are already known

Types of Assessors

Relevance
- An information intermediary
  - Search expert
  - Know the searching strategies
  - Know how to ask
- An subject specialist
  - Subject expert
  - Know the subject matter of the request

Pertinence
- The requester
  - User
  - Layman

Recall and Precision

- Recall
  - The index's ability to let relevant documents through the filter
  - A ratio of the relevant documents retrieved to the total number of relevant documents potentially available
  - Measure the completeness of the output
- Precision
  - The index's ability to hold back documents not relevant to the user
  - A ratio of the relevant documents retrieved to the number of documents retrieved
  - Measure the preciseness of the output

Consider Indexes & Abstracts ONLY

Index

- Be easy to read
- Be detailed
- Reflect the user's viewpoint
- Have multiple entry points for an idea

Abstracts

- Represent the item's aboutness
- Exclude unimportant information
- Be error free
- Be brief and readable

Overall Evaluations

- Evaluate index in terms of information retrieval evaluation
- A Controlled Environment
- Test Collection

測試集 (Test Collections)

- 組成要素
  - 文件集 (Document Set; Document Collection)
  - 查詢問題 (Query; Topic)
  - 相關判斷 (Relevant Judgement)
- 用途
  - 設計與發展: 系統測試
  - 評估: 系統效能(Effectiveness)之測量
  - 比較: 不同系統與不同技術間之比較
- 評比
  - 根據不同的目的而有不同的評比項目
  - 量化的測量準則，如Precision與Recall

測試集(Test Collections) (續)

- 小型測試集
  - 早期: Cranfield
  - 英文: SMART Collections, OHSUMED, Cystic Fibrosis, LISA....
  - 日文: BMIR-J2
- 大型評比環境: 提供測試集及研討的論壇
  - 美國: TREC
  - 日本: NTCIR, IREX,
  - 歐洲: AMARYLLIS

TREC～簡介

- TREC: Text REtrieval Conference
- 主辦: NIST及DARPA，為 TIPSTER文件計劃之子計劃之一
- 文件集
  - 5GB以上
  - 數百萬篇文件

Sample Document

<DOC>
<DOCNO>cdn_chi_19980508_0003</DOCNO>
<LANG>CH</LANG>
<HEADLINE>中共新對臺政策趨向務實彈性</HEADLINE>
<DATE>1998-05-08</DATE>
<TEXT>
<P>港媒報導，中共國家主席江澤民將在近期內於北京舉行的全國對臺工作會議中，提出帶有可操作性的新對臺政策，也就是把促進中國的完全統一，提列爲中共全黨今後五年的重大政治任務。</P>
<P>香港「星島日報」報導，由於江澤民極爲重視對臺工作，因此把推動兩岸統一的進程視爲他任內最重要的工作之一。預料在這次全國對臺工作會議中，江澤民將在「江八點」的基礎上，提出新的對臺政策，使促進兩岸統一的工作走向務實，並具有明確的可操作性。</P>
<P>江澤民也將要求中共全黨都要把促進中國統一當作全黨的中心大事來抓，因此，預計新的中央對臺工作領導小組成員、各省市自治區黨委書記、國務院臺辦、各部委臺辦、各省市區臺辦、海協會、軍方對臺部門等負責人都將參加此次全國對臺工作會議。</P>
<P>報導引述北京消息人士的話說，新的中共中央對臺工作領導小組成員已經確定：由江澤民出任組長、國務院副總理錢其琛任副組長、中共統戰部部長王兆國任秘書長；小組其他成員還包括：海協會會長汪道涵、共軍副總參謀長熊光楷、國家安全部長許永躍和中臺辦、國臺辦主任陳雲林共七人。</P>
<P>據了解，許永耀是接任已出掌公安部部長的原國家安全部長賈春旺一職，而躋身中共中央對臺工作領導小組，而另一新人則是國臺辦主任陳雲林。</P>
<P>在這項會議後，國臺辦將接著舉行兩天的全國臺辦主任會議，由國臺辦主任陳雲林主持，傳達落實全國對臺工作會議的精神，並具體部署今年的對臺工作。</P>
</TEXT>
</DOC>

Sample Topic

<TOPIC>
<NUM>001</NUM>
<SLANG>KR</SLANG>
<TLANG>CH</TLANG>
<TITLE>金融商品</TITLE>
<DESC>
介紹金融機關的金融商品
</DESC>
<NARR>
相關文件不僅紀錄了金融機關裡開發的金融商品的商品特性以及商品零售對象，而且必須包括金融機關名稱以及金融商品名稱。
</NARR>
<CONC>
銀行，金融機關，金融商品，應行商品
</CONC>
</TOPIC>

TREC～查詢主題

- 主題結構與長度
- 主題建構
- 主題篩選
  - pre-search
  - 判斷相關文件的數量

	欄位	字數（包含停字） 最小字數	最大字數	平均字數
TREC-1 (51-100)	Total	44	250	107.4
	Title	1	11	3.8
	Description	5	41	17.9
	Narrative	23	209	64.5
	Concepts	4	111	21.2
TREC-2 (101-150)	Total	54	231	130.8
	Title	2	9	4.9
	Description	6	41	18.7
	Narrative	27	165	78.8
	Concepts	3	88	28.5
TREC-3 (151-200)	Total	49	180	103.4
	Title	2	20	6.5
	Description	9	42	22.3
	Narrative	26	146	74.6
TREC-4 (201-250)	Total	8	33	16.3
	Description	8	33	16.3
TREC-5 (251-300)	Total	29	213	82.7
	Title	2	10	3.8
	Description	6	40	15.7
	Narrative	19	168	63.2
TREC-6 (301-350)	Total	47	156	88.4
	Title	1	5	2.7
	Description	5	62	20.4
	Narrative	17	142	65.3

TREC～相關判斷

- 判斷方法
  - Pooling Method
  - 人工判斷
- 判斷基準: 二元式, 相關與不相關
- 相關判斷品質
  - 完整性
  - 一致性

TREC～評比

Tasks/Tracks		TREC1	TREC2	TREC3	TREC4	TREC5	TREC6	TREC7
Main Tasks	Routing	▦	▦	▦	▦	▦	▦	
	Adhoc	▦	▦	▦	▦	▦	▦	▦
Confusion	Confusion				▦	▦		
	Spoken Document Retrieval						▦	▦
Database Merging					▦	▦		
Filtering					▦	▦	▦	▦
High Precision							▦	▦
Interactive					▦	▦	▦	▦
Multilingual	Cross Language						▦	▦
	Spanish			▦	▦	▦		
	Chinese					▦	▦	
Natural Language Processing						▦	▦	
Query								▦
Very Large Corpus							▦	▦

TREC～質疑與負面評價

- 測試集方面
  - 查詢主題
    - 並非真實的使用者需求, 過於人工化
    - 缺乏需求情境的描述
  - 相關判斷
    - 二元式的相關判斷不實際
    - pooling method會遺失相關文件, 導致回收率不準確
    - 品質與一致性
- 效益測量方面
  - 只關注量化測量
  - 回收率的問題
  - 適合作系統間的比較, 但不適合作評估

TREC～質疑與負面評價 (續)

- 評比程序方面
  - 互動式檢索
    - 缺乏使用者介入
    - 靜態的資訊需求不切實際

NTCIR ～簡介

- NTCIR: NACSIS Test Collections for IR
- 主辦: NACSIS(日本國家科學資訊系統中心)
- 發展背景
  - 大型日文標竿測試集的需求
  - 跨語言檢索的研究發展需要
- 文件集
  - 來源爲NACSIS Academic Conference Papers Database
  - 主要爲會議論文的摘要
  - 超過330,000篇文件, 其中超過1/2爲英日文對照之文件
  - 有部分包含part-of-speech tags

NTCIR～查詢主題

- 來源: 搜集真實的使用者需求, 再據其修正改寫
- 每個學科主題領域各有100個測試主題
- 組成結構

<TOPIC q=nnnn>編號
<title>標題 </title>
<description>資訊需求之簡短描述 </description>
<narrative>資訊需求之細部描述, 包括更進一步的解釋, 名詞的定義, 背景知識, 檢索的目的, 預期的相關文件數量, 希望的文件類型, 相關判斷的標準等
</narrative>
<concepts>相關概念的關鍵詞 </concepts>

NTCIR ～相關判斷

- 判斷方法
  - 利用pooling method先進行篩選
  - 由各主題專家, 及查詢主題的建構者進行判斷
- 判斷基準
  - S: 非常相關
  - A: 相關
  - B: 部分相關
  - C: 不相關

Performance Measures

- Rigid
  - S, A: relevant
  - B, C: irrelevant
- Relax
  - S, A, B: relevant
  - C: irrelevant

NTCIR～評比

- Ad-hoc Information Retrieval Task
- Cross-Language Information Retrieval Task
  - 利用日文查詢主題檢索英文文件
  - 共有21個查詢主題, 其相關判斷包括英文文件與日文文件
  - 系統可選擇自動或人工建立查詢問題
  - 系統需送回前1000篇檢索結果

Standards

- ANSI/NISO Z39.14 – 1997 Guidelines for abstracts
- ANSI/Z39.4 – 1984 Basic criteria for indexes
- ISO 999: 1996 Guidelines for the content, organization and presentation of indexes
- BS 3700: 1988 Preparing indexes to books, periodicals, and other documents
- BS 6529: 1984 Examining documents, determining their subjects and selecting indexing terms
- ANSI/NISO Z39.50 – 1995 Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (Version 3)