Indexing and Abstracting
Lecture 06 -- 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
- The history of indexing evaluation is as long as indexing
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 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

- 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

Who is Buck Rogers

BUCK ROGERS in the 25th Century aired on NBC beginning September 20, 1979 and ran until April 16, 1981. The Series almost completed two seasons with a total of 37 episodes.
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
- Pertinence is the relationship between a document and an information need
- 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
Relevance and Pertinence

(continued)

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

Types of People for Judgment

- 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
- The requester
  - Layman
  - The one to judge pertinence
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

Index Quality

- Is easy to read
- Is detailed
- Reflects the user’s viewpoint
- Has multiple entry points for an idea
Points to Take Care for Indexing

- Coverage is complete
- Consistency in term choices
- Term choices are appropriate to the nature of the users
- There is adequacy of cross-references but they are not overzealously done
- There are not numerous strings of undifferentiated subheadings
- Subheadings truly reflect the main heading
- No incorrect or missing locators
- No excessive strings of undifferentiated locators
- No proper names missing
- Alphabetization is consistent throughout the text
- Misspellings are all corrected
- No indexing the same topic under different index terms without proper cross-referencing
- No circular cross-references

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
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
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

### TREC~查詢主題

<table>
<thead>
<tr>
<th>國家</th>
<th>資料</th>
<th>資料</th>
<th>資料</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC-1</td>
<td>Title</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>TREC-1</td>
<td>Description</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>TREC-1</td>
<td>Narrative</td>
<td>23</td>
<td>209</td>
</tr>
<tr>
<td>TREC-1</td>
<td>Concepts</td>
<td>4</td>
<td>111</td>
</tr>
<tr>
<td>TREC-2</td>
<td>Title</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>TREC-2</td>
<td>Description</td>
<td>6</td>
<td>41</td>
</tr>
<tr>
<td>TREC-2</td>
<td>Narrative</td>
<td>27</td>
<td>165</td>
</tr>
<tr>
<td>TREC-2</td>
<td>Concepts</td>
<td>3</td>
<td>88</td>
</tr>
<tr>
<td>TREC-3</td>
<td>Title</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>TREC-3</td>
<td>Description</td>
<td>9</td>
<td>42</td>
</tr>
<tr>
<td>TREC-3</td>
<td>Narrative</td>
<td>26</td>
<td>146</td>
</tr>
<tr>
<td>TREC-4</td>
<td>Title</td>
<td>8</td>
<td>33</td>
</tr>
<tr>
<td>TREC-4</td>
<td>Description</td>
<td>8</td>
<td>33</td>
</tr>
<tr>
<td>TREC-4</td>
<td>Narrative</td>
<td>29</td>
<td>213</td>
</tr>
<tr>
<td>TREC-5</td>
<td>Title</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>TREC-5</td>
<td>Description</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>TREC-5</td>
<td>Narrative</td>
<td>19</td>
<td>168</td>
</tr>
<tr>
<td>TREC-6</td>
<td>Title</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>TREC-6</td>
<td>Description</td>
<td>5</td>
<td>62</td>
</tr>
<tr>
<td>TREC-6</td>
<td>Narrative</td>
<td>17</td>
<td>142</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Tasks/Tracks</th>
<th>TREC1</th>
<th>TREC2</th>
<th>TREC3</th>
<th>TREC4</th>
<th>TREC5</th>
<th>TREC6</th>
<th>TREC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Tasks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routing</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Adhoc</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Confusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confusion</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Spoken Document Retrieval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database Merging</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Filtering</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>High Precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilingual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross Language</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Spanish</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Chinese</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Query</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Large Corpus</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

### 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
  - 來源為NACSIS Academic Conference Papers Database
  - 主要為會議論文的摘要
  - 超過330,000篇文件, 其中超過1/2為英日文對照之文件
  - 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
  - A: 相關
  - B: 部分相關
  - C: 不相關
- 精確率計算：依測試項目的不同而有不同
  - Relevant quel: B與C均視為不相關
  - Partial Relevant quel: A與B均視為相關
NTCIR～評比

- Ad-hoc Information Retrieval Task
- Cross-lingual 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
- Automatic Term Extraction and Role Analysis Task
  - Automatic Term Extraction: 從題名與摘要中抽取技術用語
  - Role Analysis Task: 抽取出主題，研究方法，研究程序

Evaluation of Abstracts

- Should represent what the item is about
- Should exclude unimportant information
- Should be error free
- Should be brief and readable
Standards

- 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

Editing

- Correct alphabetizing
- Divide long multilevel headings
- Eliminate synonymous headings by consolidation under preferred term
- Recheck the editing
- Correct too many locators attached to an entry
- Correct ambiguous headings: base? Is it military, or is it a Tiffany lamp?
- Verify cross-references
- Correct spelling and capitalization
- Check punctuation and caps
- Check need to add entries
- Are main headings relevant and/or needed?
- Are locators correct?
- Is the index overdone?
- Is the index underdone?