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 gives 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 concept 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

- 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
  - 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
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

Language & Information Processing System, LIS, NTU Indexing & Abstracting Lecture07 25/42

Language & Information Processing System, LIS, NTU Indexing & Abstracting Lecture07 26/42
Sample Topic

<TOPIC>
    <NUM>001</NUM>
    <SLANG>KR</SLANG>
    <TLANG>CH</TLANG>
    <TITLE>
        金
        ም
        ൖ
    </TITLE>
    <DESC>
        ʧୗ
        金
        ም
    </DESC>
</TOPIC>

TREC～相關判斷

- 判斷方法
  - Pooling Method
  - 人工判斷

- 判斷準：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>Routing</td>
<td>Adhoc</td>
<td>Confusion</td>
<td>Spoken Document Retrieval</td>
<td>Database Merging</td>
<td>Filtering</td>
<td>High Precision</td>
</tr>
<tr>
<td></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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></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></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></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></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></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
  - 問題主題
    - 並非真實的使用者需求，過於人工化
    - 缺乏需求尋找的描述
  - 相關判斷
    - 二元式的相關判斷不實際
    - pooling method會損失相關文件，導致回收率不準確
    - 品質與一致性
- 效益測量方面
  - 只關注數字測試
  - 回收率的問題
  - 合作系統間的比較，但不進行合作評估

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
  - 大型日文語料庫測試集的需求
  - 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
  - 約330,000篇文件，其中約為1/2為英文對照之文件
  - 有部分包含part-of-speech tags

TREC～質疑與負面評價（續）

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

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
  <TOPI C q=nnnn>標題
  <title>標題</title>
  <description>資料需求之簡短描述</description>
  <narrative>資料需求之詳細描述，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</concept s>
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
  - 利用日文查詢主題索引英文文件
  - 共有21個查詢主題，其相關判斷包括英文文件
  - 系統可選擇自動或人工建立查詢問題
  - 系統書送回前1000篇检索結果

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