Indexing and Abstracting
Lecture 09 -- Automatic Indexing

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

Outline
- What's the subject indexing?
- Types of subject indexing
- The taxonomy for subject indexing
- Index Structures
- The models for automatic indexing
- 3-tier model for automatic indexing
- Natural language processing techniques
- Future trends
Subject Indexing

- The function of indexing is to describe the "aboutness" of documents
- Indexing terms are used to present content of document
- Challenge
  - how to select a set of indexing terms to represent the document contain thousands of words faithfully
- Process
  - subject analysis
  - subject translation

Process of Subject Indexing

- Subject analysis
  - Analyze content of texts and distill the subject concepts
  - Basis of subject indexing

- Subject translation
  - Translate the subject concepts to index terms
  - Two approaches
    - natural language indexing (free-text indexing)
    - controlled vocabulary indexing
Tasks of Indexing

- Analysis of the subject content of the document
- Review of authorities to understand the subject content
- Review of indexing policies and authorities to aid in the correct assignment of terms
- Presentation of the index terms in the appropriate order of the indexing system
- Weighting of index terms
- Quality control of the index terms
Indexing Consistency

- The degree of agreement in the representation of the essential information content of the document by certain sets of indexing terms selected individually and independently by each of the indexers in the group.

Indexing Consistency Rating

- all studies indicated that consistency was very low
- a figure of 30% often was used
- Indexing consistency can vary on several factors
  - familiarity with the indexing policies
  - experience with the specific subject
  - the document most recently indexed
  - the time allowed to complete the task
How to Measure Consistency

- inter-indexer consistency
  - the overlap in index term assignment by two or more indexers for the same document
- intra-indexer consistency
  - the same indexer indexes the same document at two different times

Manual Aids to Indexing

- Manual Aids
  - Vocabulary control
  - thesauri
  - scope notes
- indexer may choose not to use manual aids
  - takes additional time
  - relevance of the aid to the problem is not apparent
  - indexer believes there is no problem at all
Machine-Readable Indexing Aids

- The indexer’s tools included authority files, policy manuals, handbooks, textbooks, sheets in print form.
- Machine Readable Indexing resources are available.

Pre- versus Post-coordination

- Pre-coordinated indexing term
  - complex/compound concepts are represented in a single term
- Post-coordinated indexing term
  - concepts are joined at the time of retrieval.
Controlled versus Uncontrolled

- Controlled Indexing
  - may be selected from a hierarchical thesaurus
  - may be selected from a list of classification level subject headings
- Uncontrolled Indexing
  - natural language terms (free terms) from texts with or without standardization

Automatic versus Manual

- Automatic indexing
  - Apply computers to proceed the indexing task
- Manual indexing
  - Human indexers proceed the indexing task
The taxonomy for Indexing Scheme

- Use 3-tuple to represent possible indexing scheme
  - The first element denotes pre-coordinated (+) or post-coordinated (-)
  - The second element denotes controlled (+) or uncontrolled (-)
  - The third element denotes automatic (+) or manual (-)
- IS(-, +, +) represents post-coordinated, controlled, and automatic indexing

Automatic Indexing

- Most works are devoted to automatic free-text indexing
- Few works concern the automatic controlled-vocabulary indexing
The effectiveness of any content analysis or indexing system is controlled by two parameters:

- **Indexing exhaustivity**
  - the degree to which all aspects of the subject matter of a text item are actually recognized

- **Term specificity**
  - the degree of breadth or narrowness of the terms

**Term Specificity**

- Broad terms cannot distinguish relevant from irrelevant items
- Narrow terms retrieve relatively fewer items, but most of the retrieved materials are likely to be helpful to users
Approaches for Automatic Indexing

- Semantic Approach
  - based on understanding texts
  - domain-dependent
- Syntactic Approach
  - based on syntactic analysis of texts
  - language-dependent
- Statistical Approach
  - based on the statistics of terms
  - portable

Term Frequency

- Function words
  - for example, "and", "or", "of", "but", ...
  - *the frequencies of these words are high in all texts*
- Content words
  - words that actually relate to document content tend to occur with greatly *varying frequencies in the different texts* of a collection
  - the frequency of content word may be used to indicate term importance for content representation.
A Frequency-Based Indexing Method

- Eliminate common function words from the document texts by consulting a special dictionary, or stop list, containing a list of high frequency function words.
- Compute the term frequency $tf_{ij}$ for all remaining terms $T_j$ in each document $D_i$, specifying the number of occurrences of $T_j$ in $D_i$.
- Choose a threshold frequency $T$, and assign to each document $D_i$ all term $T_j$ for which $tf_{ij} > T$.

Document Frequency (DF)

- The number of documents which contain the designated word for a certain collection.
- $df_j = df(T_j) = \text{NumberOfDocumentContain}(T_j)$.
Compose a Single Frequency-Based Indexing Model

- Best indexing terms are those that occur frequently in individual documents but rarely in the remainder of the collection.
- A simple combined term importance indicator is

\[ w_{ij} = tf_{ij} \times \log \frac{N}{df_j} \]

An Improved Indexing Policy for Free-Term Indexing

- Eliminating common function words
- Computing the value of \( w_{ij} \) for each term \( T_j \) in each document \( D_i \)
- Assigning to the documents a collection of all terms with sufficiently high \( (tf \times idf) \) factors
Problems of Traditional Model for Control-Vocabulary Indexing

- Term statistics
  - Term frequency (TF)
  - Document frequency (DF)
  - Inverse document frequency (IDF = log(N/DF))
- Traditional model: TF×IDF
- High DF words
  - Common words
  - Domain-specific words
  - Subject-specific words

A+B+C is the words with high DF and low IDF
A = Common words
B = Domain-specific words
C = Subject-specific words

3-Tier Model for Automatic Indexing

TF of the same document
DF of the same subject
IDF of the same domain
### Basic Idea

<table>
<thead>
<tr>
<th>Subject Headings</th>
<th>Learning Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>$R_1 = {w_{i1}, w_{i2}, w_{i3}, w_{i4}, \ldots, w_{ii}}$</td>
</tr>
<tr>
<td>$H_2$</td>
<td>$R_2 = {w_{i1}, w_{i2}, w_{i3}, w_{i4}, \ldots, w_{ii}}$</td>
</tr>
<tr>
<td>$H_3$</td>
<td>$R_3 = {w_{i1}, w_{i2}, w_{i3}, w_{i4}, \ldots, w_{ii}}$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$H_j$</td>
<td>$R_j = {w_{ij1}, w_{ij2}, w_{ij3}, w_{ij4}, \ldots, w_{ij}}$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$H_m$</td>
<td>$R_m = {w_{im1}, w_{im2}, w_{im3}, w_{im4}, \ldots, w_{imm}}$</td>
</tr>
</tbody>
</table>
The Scheme for Term Weight

DF

\[ H_1 \]

\[ H_2 \]

\[ H_3 \]

\[ H_m \]

IDF

DF versus IDF

<table>
<thead>
<tr>
<th></th>
<th>DF original set</th>
<th>IDF combined set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Words</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Domain-specific Words</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Subject-specific Words</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>
Term Weighting

\[ \text{Weight} = \text{TF} \times \text{DF}_{\text{original set}} \times \text{IDF}_{\text{combined set}} \]

\[ \text{Weight}_{ik} = \text{TF}_{ik} \times \text{OSDF}_{nk} \times \text{CSIDF}_{mk} \]

- \( \text{Weight}_{ik} \) = weight of term \( k \) in document \( i \)
- \( \text{TF}_{ik} \) = frequency of term \( k \) in document \( i \)
- \( \text{OSDF}_{nk} \) = document frequency of term \( k \) in original document collection \( n \)
- \( \text{CSIDF}_{mk} \) = inverse document frequency of term \( k \) in combined document collection \( m \)

Training Stage

- Select experimental texts and controlled vocabulary
- Select testing subjects
- Train parameters for the proposed model

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Testing Set</th>
<th>(Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>40,000</td>
<td>20,000</td>
<td>60,000</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>(Total)</td>
<td>40,000</td>
<td>20,400</td>
<td>60,400</td>
</tr>
</tbody>
</table>
Testing Stage

- Compute the indexing score for testing texts

\[
\text{Indexing Score} = \frac{\sum (OSDF \times CSIDF) \times (TF)}{\text{number of words in the document}}
\]

(weight of a word = 0, when the word isn't included in \( R_j \))

- Thresholding

\[
\text{IF} \quad IS > M_j \\
\text{THEN} \\
H_j \text{ is assigned to the document}
\]

Evaluation Criteria

- Indexing precision

\[
\text{Indexing Precision} = \frac{\text{正確索引之文件數}}{\text{模型索引之文件數}}
\]

- Indexing recall

\[
\text{Indexing Recall} = \frac{\text{模型正確索引之文件數}}{\text{文件集中應索引之文件數}}
\]
### Abstract Part

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision(%)</td>
<td>Recall(%)</td>
</tr>
<tr>
<td>0.1</td>
<td>77.31</td>
<td>99.62</td>
</tr>
<tr>
<td>0.2</td>
<td>87.17</td>
<td>98.83</td>
</tr>
<tr>
<td>0.3</td>
<td>91.68</td>
<td>97.36</td>
</tr>
<tr>
<td>0.4</td>
<td>93.92</td>
<td>95.32</td>
</tr>
<tr>
<td>0.5</td>
<td>95.49</td>
<td>92.88</td>
</tr>
<tr>
<td>0.6</td>
<td>96.33</td>
<td>90.24</td>
</tr>
<tr>
<td>0.7</td>
<td>96.92</td>
<td>87.31</td>
</tr>
<tr>
<td>0.8</td>
<td>97.32</td>
<td>84.51</td>
</tr>
<tr>
<td>0.9</td>
<td>97.62</td>
<td>81.69</td>
</tr>
<tr>
<td>1.0</td>
<td>97.83</td>
<td>79.09</td>
</tr>
</tbody>
</table>

### Title Part

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision(%)</td>
<td>Recall(%)</td>
</tr>
<tr>
<td>0.1</td>
<td>80.52</td>
<td>76.48</td>
</tr>
<tr>
<td>0.2</td>
<td>86.85</td>
<td>74.86</td>
</tr>
<tr>
<td>0.3</td>
<td>89.94</td>
<td>72.86</td>
</tr>
<tr>
<td>0.4</td>
<td>91.92</td>
<td>70.73</td>
</tr>
<tr>
<td>0.5</td>
<td>92.98</td>
<td>68.50</td>
</tr>
<tr>
<td>0.6</td>
<td>93.60</td>
<td>66.09</td>
</tr>
<tr>
<td>0.7</td>
<td>94.10</td>
<td>63.78</td>
</tr>
<tr>
<td>0.8</td>
<td>94.51</td>
<td>61.47</td>
</tr>
<tr>
<td>0.9</td>
<td>94.88</td>
<td>59.30</td>
</tr>
<tr>
<td>1.0</td>
<td>95.19</td>
<td>57.12</td>
</tr>
</tbody>
</table>
### Comparison to Traditional Model

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision(%)</td>
<td>Recall(%)</td>
</tr>
<tr>
<td>0.1</td>
<td>86.67</td>
<td>97.80</td>
</tr>
<tr>
<td>0.2</td>
<td>93.33</td>
<td>89.01</td>
</tr>
<tr>
<td>0.3</td>
<td>95.08</td>
<td>73.26</td>
</tr>
<tr>
<td>0.4</td>
<td>96.10</td>
<td>54.45</td>
</tr>
<tr>
<td>0.5</td>
<td>97.09</td>
<td>38.04</td>
</tr>
<tr>
<td>0.6</td>
<td>98.53</td>
<td>24.84</td>
</tr>
<tr>
<td>0.7</td>
<td>99.81</td>
<td>15.46</td>
</tr>
<tr>
<td>0.8</td>
<td>99.89</td>
<td>9.47</td>
</tr>
<tr>
<td>0.9</td>
<td>100.00</td>
<td>5.70</td>
</tr>
<tr>
<td>1.0</td>
<td>100.00</td>
<td>3.45</td>
</tr>
</tbody>
</table>

### Related Research

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive(%)</td>
<td>Negative(%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>97.85</td>
<td>7.67</td>
</tr>
<tr>
<td>0.4</td>
<td>96.11</td>
<td>4.98</td>
</tr>
<tr>
<td>0.5</td>
<td>94.00</td>
<td>4.25</td>
</tr>
<tr>
<td>Leung &amp; Kan</td>
<td>89.70</td>
<td>4.88</td>
</tr>
</tbody>
</table>
Comparisons for Abstract

- **Training part**
  - Precision > 90%, when threshold between 0.27 and 0.61
  - Both precision and recall > 94%, when threshold = 0.43

- **Testing part**
  - Both precision and recall > 90%, when threshold = 0.27

- **Training part and testing part**
  - Recall > 96% and keep precision > 77%
  - Precision > 97% and keep recall > 71%

Comparisons for Title

- **Training Part**
  - Precision > 90% and recall > 70%, when threshold = 0.4
  - Both precision and recall > 76%, when threshold = 0.1

- **Testing Part**
  - Precision > 90% and recall > 70%, when threshold = 0.3
  - Both precision and recall > 78%, when threshold= 0.1

- **Training part and testing part**
  - Precision = 90%, we can keep the recall above 70%

- **The appropriate threshold for this application is 0.27**